

**Role of complement genetic variants in
inflammatory diseases by an interactive
database and protein structure modelling**

by

Amy Jane Osborne

A thesis presented for the degree of

Doctor of Philosophy

in the

**Research Department of Structural and Molecular
Biology,
University College London**

August 2018

Declaration

I, Amy Jane Osborne, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The rare diseases atypical haemolytic uraemic syndrome (aHUS) and C3 glomerulopathy (C3G) are associated with dysregulation of complement activation. It is unclear which genes most frequently predispose to aHUS or C3G. Accordingly, a six-centre analysis of 610 rare genetic variants in 13 mostly complement genes from >3500 patients with aHUS and C3G was performed. A new interactive Database of Complement Gene Variants was developed to extract allele frequencies for these 13 genes using the Exome Aggregation Consortium server as the reference genome. For aHUS, significantly more protein-altering rare variation was found in the five genes *CFH*, *CFI*, *CD46*, *C3* and *DGKE* than in ExAC. For C3G, an association was only found for rare variants in *C3* and the N-terminal C3b-binding or C-terminal non-surface-associated regions of factor H (FH). FH is the major regulator of C3b and its Tyr402His polymorphism is an age-related macular degeneration risk-factor. To better understand FH complement binding, the solution structures of both allotypes were studied. Starting from known FH short complement regulator domains and glycan structures, small angle X-ray scattering data were fitted using Monte Carlo methods to determine atomistic structures for monomeric FH. The analysis of 29,715 physically realistic but randomised FH conformations resulted in 100 similar best-fit FH structures for each allotype. Two distinct molecular structures resulted; an extended N-terminal domain arrangement with a folded-back C-terminus, or an extended C-terminus and folded-back N-terminus. To clarify FH functional roles in host protection, crystal structures for the FH complexes with C3b and C3dg revealed that the extended N-terminal conformation accounted for C3b fluid phase regulation, the extended C-terminal conformation accounted for C3d binding, and both conformations accounted for bivalent FH binding to the host cell-surface. Finally, statistical analyses indicated that the structural location of rare variants in complement may predict the occurrences of aHUS or C3G.

Impact Statement

The analyses and insights presented in this PhD thesis could be put to a number of beneficial uses, both outside and inside of the academic environment. The main focus of this thesis is on the molecular mechanisms of the complement inflammatory pathway, which provides rapid protection from pathogens, and its associations with genetic disease. Thus, disrupted complement regulation is associated with age-related macular degeneration (AMD), the most common cause of blindness in the West, and atypical haemolytic uraemic syndrome (aHUS) and C3 glomerulopathy (C3G), two rare and fatal diseases. The first major analysis in this thesis confirmed the association of rare variants in the five complement and related genes *CFH*, *CFI*, *C3*, *CD46* and *DGKE* with aHUS, and in *CFH* and *C3* with C3G. The reference datasets comprised unselected individuals from large-scale sequencing, such as the Exome Aggregation Consortium. Overall, my results explain how genetic changes in the same inflammatory pathway and/or proteins result in the different pathologies. This work included the development of the new interactive Database of Complement Gene Variants, which presents and analyses genetic variant data for complement diseases. By this, the database facilitates the assessment of the impacts of rare complement variants identified in patients thus being an important resource. For the clinician, this informs the diagnoses, treatment and outcome for patients. It also facilitates research-based collaboration between researchers and clinicians at the international level. This work was published in the Journal of Immunology (2018). In addition, for both the academic and the commercial pharmaceutical sectors, the rare variant analyses have potential for the development of new therapies for aHUS and C3G which will greatly benefit patients. The second major research finding in this thesis was that the complement regulator factor H (FH) exists in two distinct structural conformations that account for its functions. Thus, the extended N-terminal conformation accounted for C3b fluid phase regulation, an extended C-terminal conformation accounted for C3d binding, and both conformations accounted for bivalent FH binding to the target cell surface. This provided new insights into the molecular mechanisms of the complement proteins in inflammation, which may appear in education resources such as textbooks on human inflammation. Within academia, this research finding has been reviewed and returned for acceptance at the Journal of Biological Chemistry. In addition, this research provides new methodology for analysing the structural ensembles of flexible, glycosylated proteins and how these may be perturbed in human disease.

Publications

Osborne, A. J., Nan R., Miller A., Bhatt J., Gor J., and Perkins, S. J. (2018) Two distinct conformations of factor H regulate discrete complement-binding functions in the fluid phase and at cell surfaces. *J. Biol. Chem.* **293**, 17166-17187. Available from: PMID: 30217822.

Osborne, A. J., Breno, M., Borsa, N. G., Bu, F., Fremeaux-Bacchi, V., Gale, D. P., van den Heuvel, L. P., Kavanagh, D., Noris, M., Pinto, S., Rallapalli, P. M., Remuzzi, G., Rodriguez de Cordoba, S., Ruiz, A., Smith, R. J. H., Vieira-Martins, P., Volokhina, E., Wilson, V., Goodship, T. H. J. and Perkins, S. J. (2018) Statistical Validation of Rare Complement Variants Provides Insights into the Molecular Basis of Atypical Hemolytic Uremic Syndrome and C3 Glomerulopathy. *J. Immunol.* **200**, 2464-2478. Available from: PMID: 29500241.

Rodriguez, E., Rallapalli, P. M., **Osborne, A. J.** and Perkins, S. J. (2014) New functional and structural insights from updated mutational databases for complement factor H, factor I, membrane cofactor protein and C3. *Biosci. Rep.* **34**, 635-649. Available from: PMID: 25188723.

Presentations and abstracts

Osborne, A. J., Goodship, T. H. J. and Perkins, S. J. New functional and structural insights on complement-mediated diseases from genetic variation databases. Biochemical Society Conference 2015 - New Developments in Protein Structure Modelling for Biological and Clinical Research. 8th December 2015, Charles Darwin House, London, UK (Poster).

Osborne, A. J., Rodriguez de Cordoba, S., Fremeaux-Bacchi, V., Noris, M., Smith, R. J. H., van den Heuvel, L. P., Goodship, T. H. J. and Perkins, S. J. Database of Complement Gene Variants: a comprehensive database providing insights on function, structure and allele frequency for genetic variants identified in complement-mediated diseases. EURENomics Annual Meeting 11-13th May 2016, Paris, France (Oral).

Osborne, A. J., Rodriguez de Cordoba, S., Fremeaux-Bacchi, V., Noris, M., Smith, R. J. H., van den Heuvel, L. P., Goodship, T. H. J. and Perkins, S. J. Database of Complement Gene Variants: a comprehensive database providing insights on function, structure and allele frequency for genetic variants identified in complement-mediated diseases. XXVIth International Complement Workshop, 4-8th September 2016, Kanazawa, Japan. (Poster). **Winner of poster prize.**

Osborne, A. J., Rodriguez de Cordoba, S., Fremeaux-Bacchi, V., Noris, M., Smith, R. J. H., van den Heuvel, L. P., Goodship, T. H. J. and Perkins, S. J. Database of Complement Gene Variants: statistical validation of rare variants causative for atypical haemolytic uraemic syndrome and C3 glomerulopathy. aHUS Global Genetics Roundtable, 30th November 2016, Radisson Blu, London. (Oral).

Supervisory Panel

Supervisors:

Prof Stephen J. Perkins

Research Department of Structural and Molecular Biology

University College London

Prof Timothy H. J. Goodship

Institute of Genetic Medicine

Newcastle University

Dr Daniel P. Gale

Centre for Nephrology, Royal Free Hospital

University College London

Thesis Chair:

Prof Christopher W. M. Kay

Research Department of Structural and Molecular Biology

University College London

Acknowledgements

I would firstly like to thank my first supervisor Prof Steve Perkins for his continued support, guidance and encouragement throughout my PhD. I am especially grateful for his guidance in preparing manuscripts for journals and presenting at conferences, and his inspiring enthusiasm for complement factor H. Secondly, I would like to thank my second (external) supervisor Prof Tim Goodship for his support, guidance and encouragement throughout my PhD, and for the inspiring discussions on the complement-based genetics of aHUS and C3G. Thirdly, I would like to thank my second (internal) supervisor Dr Danny Gale for his support and ideas, especially for the complement gene analyses and improving the Database of Complement Gene Variants for clinical use.

I would like to thank my thesis chair, Prof Chris Kay, and also Prof Snezana Djordjevic, for their advice and encouragement throughout my PhD.

I thank Complement UK and Alexion Pharmaceuticals for the financial support they generously provided. For the Database of Complement Gene Variants, I would like to thank my collaborators for all of their inputs and useful discussions. I would like to thank Dr Pavi Rallapalli for sharing her expertise on bioinformatics and databases and for her overall guidance during my PhD. I am very grateful to EAHAD for letting me become part of their committee meetings in order to learn about mutation databases. I would like to thank Dr Jayesh Bhatt for sharing his expertise on molecular dynamics and computing, and for many useful and insightful discussions.

I would like to thank the members of the Structural Immunology group I have worked with over the past four years for their continued support, kindness and friendship, in particular Dr Nilufar Kadkhodayi-Kholghi, Dr Lindsay McDermott, Dr Jayesh Bhatt, Dr Orla Dunne, Dr Pavi Rallapalli, Valentina Spiteri, Gar Kay-Hui, Hina Iqbal and Henna Zahid.

Last but not least, I am eternally grateful to all of my family and friends who have wholeheartedly supported me throughout this PhD. Thank you.

Contents

Declaration	ii
Abstract	iii
Impact Statement	iv
Publications	v
Presentations and abstracts	v
Supervisory Panel	vi
Acknowledgements	vii
Contents	viii
List of Figures	xii
List of Tables	xiv
List of Abbreviations	xv
List of Units	xvii
Chapter 1 The complement system	1
1.1 The immune system	2
1.2 The complement system	4
1.2.1 Activation of C3	5
1.2.2 Opsonisation by C3	8
1.3 The classical pathway	9
1.4 The lectin pathway	9
1.5 The alternative pathway	10
1.6 Regulation of complement	11
1.6.1 Complement factor H	18
1.7 Diseases of complement	22
1.7.1 Atypical haemolytic uraemic syndrome	24
1.7.2 C3 glomerulopathy	29
1.7.3 Age-related macular degeneration	30
Chapter 2 Genetic variation	32
2.1 From DNA to proteins	33
2.2 Genetic inheritance	34
2.3 Genetic variation and evolution	37
2.3.1 Allele frequency	38
2.3.2 Natural selection, genetic drift and gene flow	41
2.3.3 Genetic recombination	43
2.3.4 Why do DNA mutations occur?	43
2.3.4.1 Substitution mutation	43
2.3.4.2 Other types of mutation	45
2.4 How mutations affect proteins	49
2.4.1 Non-coding mutations	52
2.4.2 Residue conservation	53
2.4.3 Loss and gain of function mutation	53
2.5 In silico predictive tools	54
2.6 Terminology of mutations, polymorphisms and variants	55

2.7	Sequencing methods	56
2.7.1	<i>Sequence databases and tools</i>	58
2.8	Genetic variants and disease	59
2.8.1	<i>Family-based studies</i>	62
2.8.2	<i>Population-based case-control studies</i>	63
2.8.3	<i>Allele frequency data for rare variants</i>	64
2.9	Methods: rare variant assessment guidelines for rare genetic diseases	67
2.10	Methods: rare variant burden tests	68
2.11	Overview of aHUS, C3G and AMD genetics	71
2.12	Methods: genetic variant web-databases	73
2.12.1	<i>Genetic variant databases for aHUS and C3G</i>	76
Chapter 3	Protein structure and dynamics	79
3.1	Protein structure and function	80
3.1.1	<i>Protein function</i>	81
3.1.2	<i>Formation of protein structure</i>	82
3.1.3	<i>Protein stability</i>	84
3.1.4	<i>Protein dynamics</i>	88
3.2	Methods for protein structure determination	89
3.2.1	<i>X-ray diffraction and scattering</i>	90
3.2.2	<i>Small angle x-ray scattering</i>	91
3.2.3	<i>Guinier analyses for small angle scattering</i>	95
3.2.4	<i>Other methods</i>	96
3.3	Atomistic molecular modelling	97
3.3.1	<i>Homology modelling</i>	98
3.3.2	<i>Ab initio modelling</i>	99
3.3.3	<i>Molecular dynamics</i>	99
3.3.4	<i>Energy minimisation</i>	102
3.3.5	<i>Monte Carlo methods</i>	103
3.3.6	<i>The SASSIE workflow</i>	104
3.3.7	<i>Methods for analysing conformational ensembles</i>	107
3.3.8	<i>Application to the complement proteins</i>	108
Chapter 4	Statistical validation of rare complement variants provides insights on the molecular basis of atypical haemolytic uraemic syndrome and C3 glomerulopathy	113
4.1	Summary	114
4.2	Introduction	115
4.3	Methods	117
4.3.1	<i>Data collection</i>	117
4.3.2	<i>Data cleansing, duplications and maintenance</i>	119
4.3.3	<i>Web-database development</i>	119
4.3.4	<i>Data retrieval</i>	120
4.3.5	<i>Rare variant burden</i>	120
4.3.6	<i>Rare variant assessment</i>	121
4.3.7	<i>Spatial distribution of missense rare variants in the proteins</i>	122
4.3.8	<i>Statistical analyses</i>	122

4.3.9	<i>Allele frequency analyses</i>	123
4.4	Results	124
4.4.1	<i>Genetic variants in aHUS and C3G</i>	124
4.4.2	<i>Rare variant frequencies in cases</i>	127
4.4.3	<i>Rare variant profiles of cases</i>	130
4.4.4	<i>Gender analyses</i>	130
4.4.5	<i>Rare variant pathogenicity classification</i>	134
4.4.6	<i>Rare variant abundance in genes</i>	134
4.4.7	<i>Gene-based rare variant burden</i>	136
4.4.8	<i>Distribution of aHUS and C3G rare variants in FH</i>	138
4.4.9	<i>Distribution of aHUS and C3G rare variants in C3</i>	141
4.4.10	<i>Distribution of aHUS and C3G rare variants in FI, CD46, FB and others</i>	141
4.4.11	<i>Minor allele frequency analyses</i>	142
4.5	Discussion	142
4.5.1	<i>Summary of rare variants in aHUS and C3G</i>	142
4.5.2	<i>Differences between aHUS and C3G using allele frequency analyses</i>	142
4.5.3	<i>Rare variant burden testing</i>	144
4.5.4	<i>Hotspots for missense rare variants</i>	145
4.5.5	<i>Utility of the Database</i>	147
Chapter 5	Two distinct conformations of factor H regulate discrete complement-binding functions in the fluid phase and at cell surfaces	148
5.1	Summary	149
5.2	Introduction	150
5.3	Methods	153
5.3.1	<i>Homology modelling of SCR-9, SCR-14 and SCR-17</i>	153
5.3.2	<i>Building of initial FH model with eight glycans</i>	155
5.3.3	<i>Building of FH model library using Monte Carlo</i>	157
5.3.4	<i>Fitting FH models to experimental scattering data</i>	158
5.3.5	<i>Parameterisation of the FH domain arrangement</i>	160
5.3.6	<i>Principal component analyses (PCA)</i>	160
5.3.7	<i>Theoretical sedimentation co-efficient values</i>	161
5.3.8	<i>Links to web servers and tools</i>	161
5.4	Results	162
5.4.1	<i>FH models for the Tyr402 and His402 allotypes</i>	162
5.4.2	<i>The NT-COM and CT-COM separation distributions</i>	164
5.4.3	<i>Best-fit conformations for FH Tyr402 and His402</i>	168
5.5	Discussion	173
5.5.1	<i>Self-association of FH</i>	175
5.5.2	<i>Atomistic modelling of two FH conformations</i>	175
5.5.3	<i>Biological significance of the FH Tyr402 and His402 models</i>	176
Chapter 6	Structural analyses rationalise the distribution of aHUS and C3G rare variants in complement factor H	180

6.1	Summary	181
6.2	Introduction	182
6.3	Methods	184
6.3.1	<i>Rare variant distributions in FH, C3, FI and FB complexes</i>	184
6.3.2	<i>Consensus SCR domain analyses</i>	184
6.3.3	<i>Surface-accessibility analyses of FH residues</i>	185
6.3.4	<i>Frequencies of rare variants in structural disulphide bridges</i>	186
6.4	Results	187
6.4.1	<i>Frequency of aHUS and C3G rare variants in hypervariable loop</i>	187
6.4.2	<i>Distributions of aHUS and C3G rare variants in complement complexes</i>	192
6.4.3	<i>Structural locations of aHUS and C3G rare variants in FH</i>	197
6.4.4	<i>Frequency of aHUS and C3G rare variants in protein disulphides</i>	199
6.5	Discussion	201
6.5.1	<i>Consensus SCR domain updated for aHUS and C3G</i>	202
6.5.2	<i>aHUS variants cluster in regulatory FH-C3b regions</i>	203
6.5.3	<i>Structural locations of FH variants differ between aHUS and C3G</i>	204
6.5.4	<i>Disulphide bond disruption is associated with aHUS and C3G</i>	205
Chapter 7	Conclusions – new findings on complement genetics and protein structures	207
7.1	Prologue	208
7.1.1	<i>Overview of complement function and the role of FH</i>	208
7.1.2	<i>Complement genetic variants in disease</i>	208
7.2	Statistical validation of rare complement variants provides insights on the molecular basis of atypical haemolytic uraemic syndrome and C3 glomerulopathy	210
7.2.1	<i>The Database of Complement Gene Variants and future work</i>	211
7.3	Two distinct conformations of factor H regulate discrete complement-binding functions in the fluid phase and at cell surfaces	216
7.3.1	<i>Stoichiometry of FH and C3b complexes and future FH work</i>	218
7.4	Structural analyses rationalise the distribution of aHUS and C3G rare variants in complement factor H	220
7.4.1	<i>Future work</i>	222
References		223
Appendix I		256
Appendix II		258
Appendix III		259
Appendix IV		260

List of Figures

Figure 1.1	Overview of the complement system	6
Figure 1.2	The activation of C3	7
Figure 1.3	The inactivation of C3b	12
Figure 1.4	Structure of the complement factor I gene and protein	14
Figure 1.5	Schematic representation of the complement factor H gene cluster and protein family	19
Figure 1.6	Morphology of the kidney glomerulus	25
Figure 1.7	Models for the pathophysiology of aHUS and C3G in the kidney glomerulus	26
Figure 1.8	Age-related macular degeneration	31
Figure 2.1	The central dogma of biology for a eukaryotic gene	35
Figure 2.2	The Hardy-Weinberg equilibrium for two autosomal alleles	40
Figure 2.3	Punnett square showing the Hardy-Weinberg law for loci with three and four autosomal alleles in a population	42
Figure 2.4	Spectrum of disease allele effects	61
Figure 3.1	The folding funnel energy landscape for a globular protein	86
Figure 3.2	A schematic representation of a small angle scattering experiment	93
Figure 3.3	The SASSIE workflow	106
Figure 3.4	Structural arrangement of two short complement regulator domains of complement factor H	111
Figure 4.1	Stacked bar analyses showing the reference AF of the variants identified in the (A) aHUS and (B) C3G datasets	125
Figure 4.2	Summary of cases and variants in aHUS and C3G	128
Figure 4.3	RV effects and classifications in aHUS and C3G	135
Figure 4.4	The RV burden (%) per gene for the nine relevant genes in the four aHUS (Allele number (AN): 634-6256), ExAC (AN: 74194-121246), EVS (AN: 8202-13005) and C3G (AN: 208-886) datasets	137
Figure 4.5	The distribution and disease allele frequencies (AFs) of non-benign missense RVs in the domains of FH, C3, FI, MCP, and FB in the aHUS and C3G datasets	139

Figure 5.1	Cartoon of the 20 SCR domains in FH	151
Figure 5.2	Atomistic modelling searches for the FH solution structure	163
Figure 5.3	The centre-of-mass separation frequencies in the FH Tyr402 and FH His402 models	165
Figure 5.4	The separation densities in the FH Tyr402 and FH His402 models	167
Figure 5.5	Principal component analyses of the 100 best-fit models for FH Tyr402 and FH His402	169
Figure 5.6	Scattering curve fits for the centroid PCA models for FH	171
Figure 5.7	Normalised Kratky plot for the experimental and best-fit FH curves	174
Figure 5.8	Best-fit FH centroid models superimposed onto C3b and C3dg	177
Figure 6.1	Location of aHUS and C3G rare missense variants on a SCR consensus sequence	188
Figure 6.2	Rare missense variants mapped onto models for FH in complex with C3b and FI	193
Figure 6.3	Rare missense variants mapped onto the crystal model for the C3 convertase	196
Figure 7.1	Screenshot of the Database of Complement Gene Variants homepage	212
Figure 7.2	Screenshot of the Advanced Search tools in the Database of Complement Gene Variants	213
Figure 7.3	Screenshot of the Search Results web-page in the Database of Complement Gene Variants	214
Figure 7.4	Screenshot of the In-Depth Variant Analysis web-page in the Database of Complement Gene Variants	215

List of Tables

Table 2.1	The genetic code	36
Table 2.2	Non-structural variant annotation	46
Table 2.3	Structural variant annotation in dbVar	48
Table 2.4	The 20 amino acids and their properties	50
Table 2.5	Cohorts represented in the ExAC data	66
Table 2.6	Populations represented in the ExAC data	66
Table 2.7	Categorisation of the genetic variants for aHUS and C3G	69
Table 2.8	PHP and SQL commands for a web-database application	75
Table 3.1	Complement and related protein domains	109
Table 3.2	Available molecular structures for the complement proteins	112
Table 4.1	Summary of mRNA and protein identifiers for the 13 genes	118
Table 4.2	The 14 common genetic variants identified in at least one of the three reference datasets (1000GP, EVS and ExAC) at an AF of $\geq 1\%$	126
Table 4.3	The total number of aHUS and C3G cases screened per gene	129
Table 4.4	Homozygous RVs present in aHUS and C3G	131
Table 4.5	Demographics of the 1231 aHUS and 116 C3G cases showing an identified RV	132
Table 4.6	Gender of the 1231 aHUS and 116 C3G cases with an identified RV	133
Table 5.1	Sources of molecular SCR structures in FH	154
Table 5.2	Homology modelling of the SCR-9, SCR-14, and SCR-17 domains	156
Table 5.3	Experimental X-ray and analytical ultracentrifugation data for FH Tyr402/Val62 and His402/Val62 and their modelling fits	159
Table 6.1	Frequencies of aHUS and C3G affected residues in the consensus SCR	191
Table 6.2	Allele frequencies of aHUS affected residues in the consensus SCR	191
Table 6.3	FH residue surface-accessibility analyses for aHUS and C3G	198
Table 6.4	The frequency of rare variation in Cys residues that affected protein disulphide bonds for aHUS, C3G and ExAC	200

List of Abbreviations

1000GP	The 1000 Genomes Project
1, 2 or 3D	one, two or three-dimensional
A	adenine
AF	allele frequency
aHUS	atypical haemolytic uraemic syndrome
AMD	age-related macular degeneration
AP	alternative pathway
AUC	analytical ultracentrifugation
BLAST	Basic Local Alignment Search Tool
C	cytosine
C1-Inh	C1 inhibitor
C3G	C3 glomerulopathy
C4bp; <i>C4bp</i>	C4 binding protein; <i>C4 binding protein</i>
CATH	Class, Architecture, Topology/fold, Homologous superfamily
<i>CD46</i>	<i>cluster of differentiation 46</i>
<i>CFB</i>	<i>complement factor B</i>
<i>CFH</i>	<i>complement factor H</i>
<i>CFHR</i>	<i>complement factor H related</i>
<i>CFI</i>	<i>complement factor I</i>
<i>CFP</i>	<i>complement factor properdin</i>
CR1; <i>CR1</i>	complement receptor type 1; <i>complement receptor type 1</i>
cryo-EM	cryo-electron microscopy
DAA	decay-accelerating activity
DAF	decay-accelerating factor
dbVar	database of genomic structural variation
DGKE; <i>DGKE</i>	diacylglycerol kinase epsilon; <i>diacylglycerol kinase epsilon</i>
D_{max}	maximum particle diameter
DSSP	dictionary of protein secondary structure
DNA	deoxyribonucleic acid
EM	electron microscopy
EVS	Exome Variant Server
ExAC	Exome Aggregation Consortium
FB	factor B
FD	factor D
FH	factor H
FHL-1	factor H-like 1
FHR	factor H related
FI	factor I
G	guanine
GoF	gain-of-function
GWAS	genome-wide association study
HS	heparan sulphate
HTML	Hypertext Markup Language
HUS	haemolytic uraemic syndrome
HWE	Hardy-Weinberg equilibrium
iC3b	inactive C3b
LDLr1/2	low-density lipoprotein receptor 1/2
LoF	loss-of-function
MAC	membrane attack complex

MAF	minor allele frequency
MASP	mannose-binding lectin-associated serine protease
MBL	mannose-binding lectin
MCP	membrane cofactor protein
MC	Monte Carlo
MD	molecular dynamics
MG	macroglobulin
MPGN	membranoproliferative glomerulonephritis
mRNA	messenger ribonucleic acid
NAMD	Nanoscale Molecular Dynamics
NCBI	National Center for Biotechnology Information
NF κ B	nuclear factor kappa-light-chain-enhancer of activated B-cells
NMR	nuclear magnetic resonance
P or FP	properdin or factor properdin
PCA	principal component analysis
PCR	polymerase chain reaction
PDB	Protein Data Bank
PHP	PHP: Hypertext Preprocessor
PISA	Protein Interfaces, Surfaces and Assemblies (software)
PLG; <i>PLG</i>	plasminogen; <i>plasminogen</i>
PSF	protein structure file
RCA	regulators of complement activation
R_G	radius of gyration
R_{XS}	mean cross-sectional radius of gyration
RNA	ribonucleic acid
RPE	retinal pigment epithelium
RV	rare variant
SAS	small-angle scattering
SAXS	small-angle X-ray scattering
SCR	short complement regulator
SQL	Structured Query Language
SIFT	'Sorting Intolerant From Tolerant'
SNP	single nucleotide polymorphism
SPR	surface plasmon resonance
T	thymine
TED	thioester domain
THBD; <i>THBD</i>	thrombomodulin; <i>thrombomodulin</i>
UTR	untranslated region
VMD	Visual Molecular Dynamics

List of Units

°	degree
°C	degree Celsius
Å	Angstrom
e	electron
J K ⁻¹ mol ⁻¹	Joule per Kelvin per mole
K	Kelvin
kDa	kilo Dalton
L	litre
M	molar (mol/L)
mg	milligram
ml	millilitre
μM	micromolar
mM	millimolar
ms	millisecond
mol	mole
nm	nanometre

Chapter One

The complement system

This introduction chapter describes the human complement system of immunity and its involvement in the two rare renal diseases atypical haemolytic syndrome and C3 glomerulopathy, and the common eye disease age-related macular degeneration. More specifically, this involves the genes, proteins and molecular interactions of the complement alternative pathway. This theory is required for the analyses presented in my three results [Chapters 4, 5 and 6](#) on genetic variants and molecular structures of the complement proteins and their roles in these diseases.

1.1 The immune system

The immune system protects the host from disease using many different biochemical components and processes. The current germ theory of disease, which was first proposed by Pasteur in 1861 and later confirmed in 1876 by Koch ([Smith, 2012](#)), attributes the onset of infectious disease to pathogens, also known as ‘germs’. Pathogens can take the form of either viruses, bacteria, parasitic worms, fungi, algae, prions or compromised host cells and their fragments. In order to remove such disease-causing pathogens, the immune system must first distinguish them from host cells. The immune system is categorised into either the innate or adaptive systems. For the innate system, a limited number of germline-encoded pattern-recognition receptors recognise conserved patterns on pathogens ([Akira et al., 2006](#)), which rapidly triggers an inflammatory response. In contrast, the adaptive system is highly specific to a particular pathogen. This specificity of the adaptive system is due to a process known as somatic hypermutation which creates an immunological memory that ultimately leads to an enhanced immune response ([Chaplin, 2010](#)). This is exploited in vaccinations against disease. For most organisms, an innate system is sufficient for survival. Only vertebrates appear to have evolved an adaptive system ([Beutler, 2004](#)). The innate system thus evolved long before the adaptive system. One of the oldest host defence systems with similar functions to the innate system is the activation of nuclear factor kappa-light-chain-enhancer of activated B-cells (NFκB) transcription factor by the Toll pathway. This system is found in the fruit fly and possibly plants. In the Toll pathway, host cells with pattern-recognition receptors known as Toll-like receptors bind to pathogens and activate NFκB, which leads to the regulation of the immune response. In addition, phagocytic cell types of the innate system may have originated from unicellular amoeba-like early eukaryotes ([Janeway, 2001](#)). In contrast, adaptive immunity is believed to have arisen in the first jawed vertebrates

(gnathostomes), however jawless fish were recently discovered to have similar lymphoid cell-based systems of adaptive immunity ([Flajnik & Kasahara, 2010](#)).

The first major line of defence is epithelial skin cells which form a physical barrier against non-host material. If pathogens enter the host, the innate system mediates an inflammatory response. By this, both humoral (fluid) and cell-based components identify and destroy the invading pathogen. For the cell-based components, this involves white blood cells or leukocytes such as mast cells, natural killer cells and basophils, and phagocytes such as macrophages, neutrophils and dendritic cells. For pathogen identification, some of these cells express Toll-like receptors which bind to pathogen-associated molecular patterns and damage-associated molecular patterns on pathogens and compromised host cells, respectively. Once identified, the innate cells eliminate pathogens via a diverse set of processes such as digestion by macrophages and the release of toxic substances and proteins by basophils and neutrophils.

In addition to innate cells, pathogens can be destroyed via the activation of the adaptive system by antigen-presenting cells such as macrophages and dendritic cells. Thus, antigen-presenting cells digest the pathogens and present their fragments as antigens to adaptive system proteins. For the adaptive system, the lymphocytes known as B-cells and T-cells are essential for mediating each of the humoral and cell-mediated immune responses, respectively. For antigen specificity, both B- and T-cells possess genetically rearranged and highly diverse antigen receptors. These receptors are sensitive enough to be able to distinguish between two antigens that differ by only a single amino acid ([Alberts, 2002](#)). For B-cell activation, the antigens on antigen-presenting cells are bound. Once activated, B cells differentiate into either plasma or memory cells. For differentiated plasma B-cells, secreted antibody immunoglobulins (Igs) bind to specific pathogens that correspond to the presented antigens. Antibodies inactivate bound pathogens by neutralisation, agglutination, precipitation and complement activation. For memory B-cells, the antigen is biochemically remembered, which speeds up the immune response for subsequent exposures. For T-cell activation, the detection of an antigen presented on a host cell causes a cell-mediated reaction. For example, host cells infected with a virus are eliminated by T-cells before the virus can replicate. In cell-mediated immunity, T-cell activation causes the development of cytotoxic T-cells and engagement of accessory cells such as macrophages. T-cells are also involved in antibody based immunity. Thus, the activation of T cells increases both the antibody response via T-

helper cells and the production of antibody by B-cells ([Pross, 2007](#)). For lymphocyte development in mammals, T-cells develop in the thymus whereas B-cells mature in the bone marrow. However, for B-cells, their name originates from the Bursa of Fabricius organ, which is the location of B-cell maturation in birds ([Katze et al., 2016](#)).

1.2 The complement system

The complement system is an important effector component of both the innate and adaptive immune systems ([Chaplin, 2010](#)). Complement consists of both soluble and membrane bound proteins and operates in plasma, on tissues, on cell surfaces and within cells ([Merle et al., 2015b](#)). Liver hepatocytes are the major source of plasma complement components, with some exceptions. However, the penetration of complement components into tissues can be limited by their large sizes. Thus, for compensation, complement components are also produced locally ([Morgan & Gasque, 1997](#)). When the complement system is activated, a proteolytic cascade leads to the opsonisation and clearance of pathogens and immune complexes, as well as B- and T-cell modulation. Complement is considered to be evolutionarily ancient. Thus, genomic analyses have identified a primitive version that consists of two complement genes corresponding to ancient C3 and complement factor B (*CFB*) which emerged in the common ancestor of Cnidaria and Bilateralia more than 1,300 million years ago ([Nonaka & Kimura, 2006](#)). A complement C3-like protein may have been present in organisms that emerged prior to sea sponges ([Elvington et al., 2016](#)). Complement was first discovered in 1891 as a heat-labile serum factor that was able to kill bacteria ([Buchner, 1891](#)). This bactericidal activity was also found to be dependent on a heat-stable factor which is now known as antibody ([Morgan, 1990](#)). Later, complement was described as a heat-labile antimicrobial component which combined with antibody to form an enzyme that killed pathogens, as part of a theory by Paul Ehrlich. Next, in 1907, Ferrata and Brand separated complement into two fractions which are now known as the complement proteins C1 and C2. This demonstrated that complement was not a single substance. After this, work by Ueno, Mayer and colleagues showed that the sequence of the classical pathway could be unravelled in reconstitution assays by adding partially purified complement components to antibody-sensitised sheep erythrocytes. Next, in the 1960s, for the classical complement pathway, Nilsson, Muller-Eberhard and colleagues isolated and characterised the components and discovered their sequence. With the development of ion-exchange chromatography and ultracentrifugation methods, this work led to the discovery of eleven distinct plasma

proteins essential for complement activation (Law & Reid, 1995; Nelson et al., 1966). In 1968, in order to reflect the order of activation, the terminology of these complement components was adjusted. A second, antibody-independent activation pathway, now known as the alternative pathway (AP), was proposed in the 1950s by Pillemer but accepted in 1970 (Law & Reid, 1995; Nesargikar et al., 2012). A third complement activation pathway, known as the lectin pathway, was discovered in 1987 by Ikeda and colleagues. For the lectin pathway, activator proteins such as mannose-binding lectin (MBL) were identified firstly in rabbit liver and serum followed by their role in child immunodeficiencies (Nesargikar et al., 2012). Complement is now known to be activated via three distinct pathways, which merge at the activation of the complement protein C3 (Figure 1.1). The classical pathway is activated when an antibody-antigen complex is formed. The lectin pathway is activated via complex carbohydrates such as bacterial membrane polysaccharides. In contrast, the AP is activated not by pathogens but by a wide variety of compounds and surfaces such as small nucleophiles, water molecules, serum proteases and perturbations. All of the protein complement components are glycoproteins. In order to protect host cells from complement attack, each of the three complement pathways are regulated.

1.2.1 Activation of C3

In order for the complement system to destroy invading pathogens, complement C3 is activated on their surfaces. This key step is common to the three complement pathways (Figure 1.1). C3 is a large protein with a molecular weight of 185 kDa that is present in plasma at a concentration of approximately 1 mg/ml (5.3 μ M) (Law & Reid, 1995). C3 is encoded by the *C3* gene on chromosome 19 in the human genome. In the synthesis of C3, four Arginine residues are removed which splits the C3 polypeptide into a 110 kDa α -chain and a 75 kDa β -chain (Law & Reid, 1995). In mature C3, the α - and β -chains are linked by disulphide bonds and noncovalent forces (Bokisch et al., 1975). For C3, in order to transform this from an inert circulating protein with an internal thioester, accessible only to small nucleophiles, into an active form capable of binding to targets, a major conformational change is required (Law & Dodds, 1997). Thus, for the activation of C3, a single peptide bond in the α -chain is cleaved, which dissociates the first 77 amino acids, known as the anaphylatoxin C3a fragment, from the C3b fragment. In C3b, the previously buried thioester is exposed, which is extremely reactive with nucleophiles such as hydroxyl and amino groups found on surfaces (Figure 1.2).

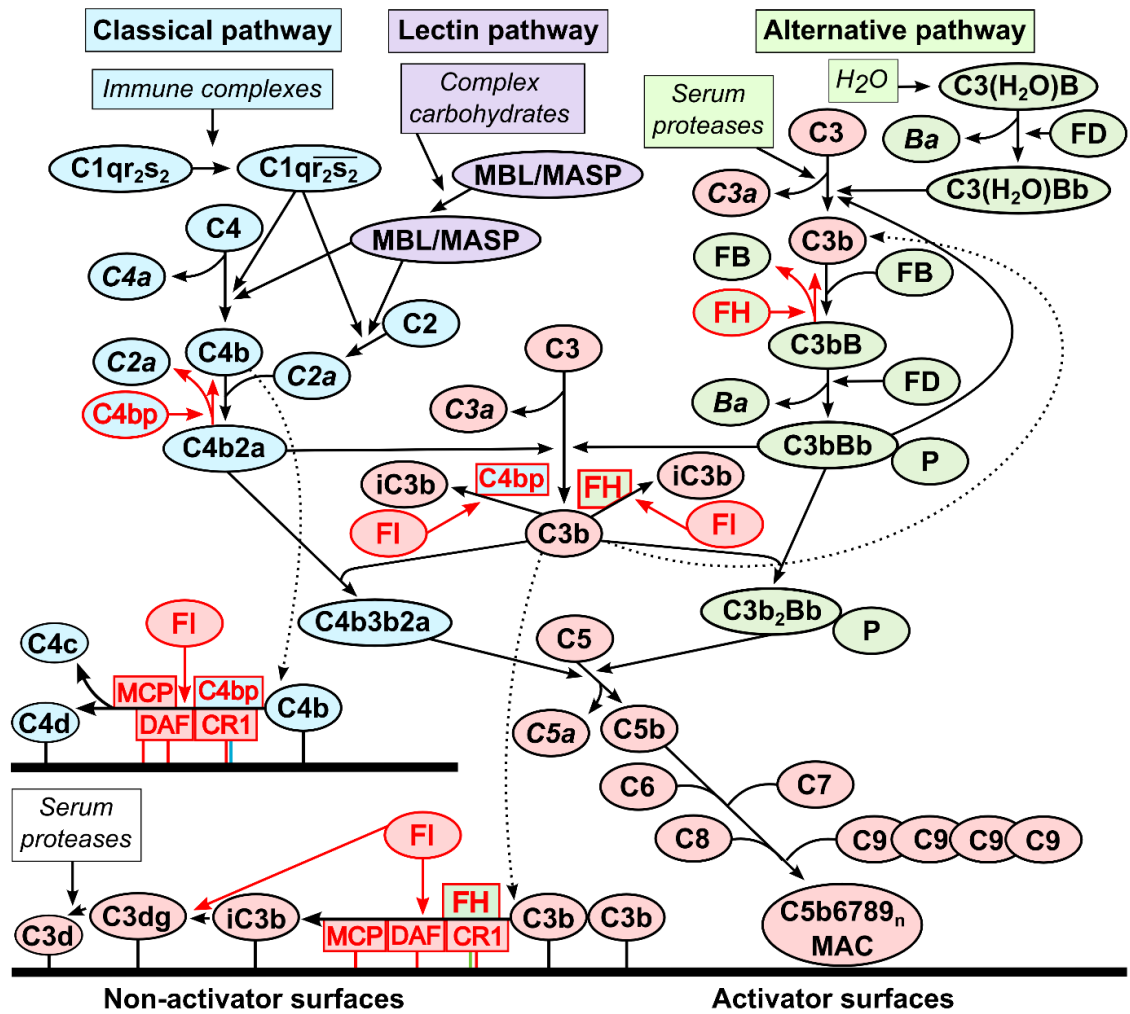


Figure 1.1 Overview of the complement system. The proteolytic activation of C3, which generates active C3b, is common to each of the classical, lectin and alternative pathways (AP). On activator surfaces, C3b binds to the C3 convertase to form a C5 convertase. This leads to formation of the membrane attack complex (MAC). The AP convertases C3bBb and C3b₂Bb are stabilised by properdin (P). Host surfaces are protected from complement attack by factor I (FI), which inactivates C3b to form iC3b in the presence of a cofactor. For FI, cofactor activity is carried out by each of C4 binding protein (C4bp), complement receptor 1 (CR1), decay-accelerating factor (DAF), factor H (FH) and membrane cofactor protein (MCP). In the fluid phase, FI regulates C3b and C4b by using FH/C4bp and C4bp as cofactors, respectively. For both C4bp and FH, decay-acceleration activities regulate the C3 and C5 convertases of the classical and the AP, respectively. The AP serves as an amplification loop for the complement system thus C3b.

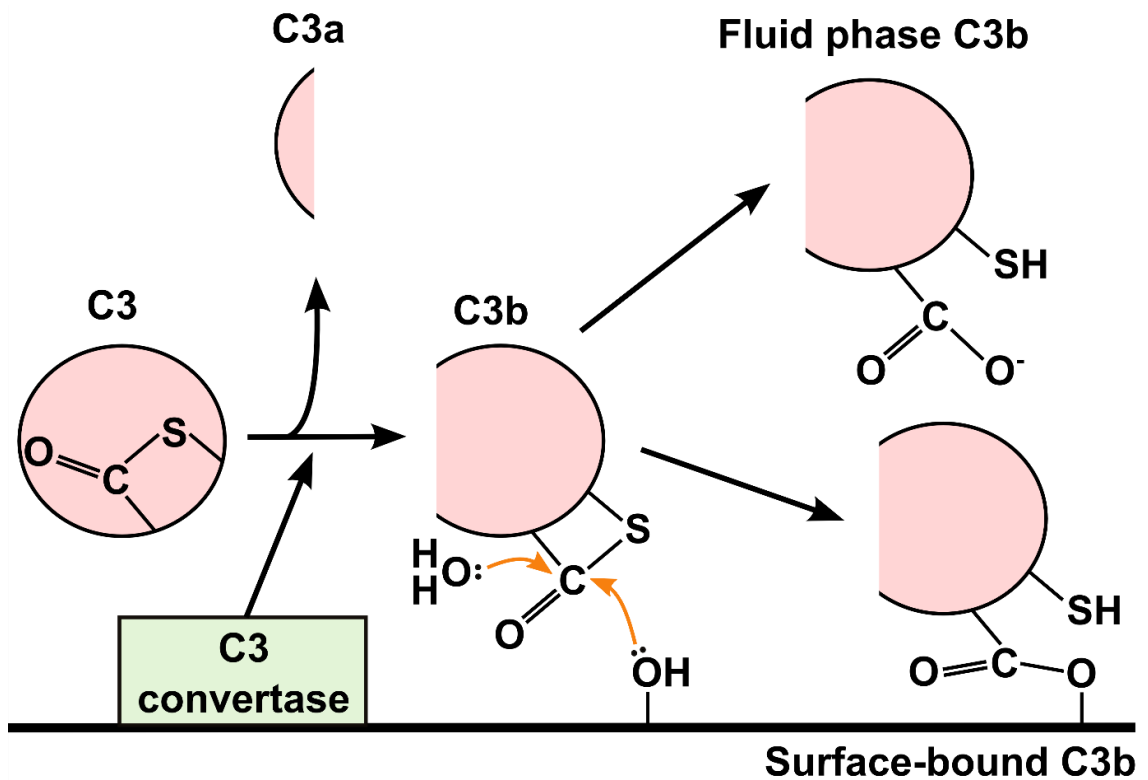


Figure 1.2 The activation of C3. In each of the three complement pathways, C3 (left) is proteolytically activated to C3b by a C3 convertase. The anaphylatoxin C3a fragment is also generated. In C3b (centre), a highly reactive thioester in a Cys residue is exposed which binds to nucleophiles such as water and hydroxyl and amine (NH) groups on surfaces. Thus C3b either reacts with water to form fluid phase C3b (top right) or covalently binds to cell surfaces via either an ester or an amide (NH) bond (bottom right).

However, for the thioester in plasma, its hydrolysis of C3 by water molecules into C3(H₂O) readily occurs due to the high concentration of water at 55 M (Law & Reid, 1995). Once C3 is cleaved to form C3b, the thioester reacts within ~60 μs, in which C3b can diffuse up to 30 nm before it becomes covalently bound to a target (Sim et al., 1981). Thus, before its hydrolysis by water, the reaction of C3b with target surfaces via the thioester is spatially restricted to local surfaces. For the proteolytic activation of C3 to C3b, catalysis by the C3 convertase serine protease is required. For each of the classical and lectin pathways, the C3 convertase complex is C4b2a. In contrast, for the AP, the C3 convertase complex is either C3bBb or C3(H₂O)Bb. In addition, instead of using its thioester, C3b can bind to the cell surface by interacting with surface molecules that serve as platforms for C3b recruitment (Merle et al., 2015a).

1.2.2 Opsonisation by C3

By opsonising target surfaces with C3b, complement triggers the destruction of the pathogen in three major ways. Firstly, it marks the target for destruction. For example, C3b molecules on activator surfaces are bound by phagocytes, including antigen-presenting cells, via their surface-expressed complement receptor 1 (CR1). The phagocytes either engulf the target or enhance its antigenic presentation to T cells. Secondly, for viral pathogens, C3b neutralises the virus by blocking their receptor-mediated entry to host cells. Thirdly, for immune complexes, C3b eases their clearance from the circulation by solubilisation. Thus, for C3b-coated immune complexes, erythrocytes bind to C3b via CR1 and transport them to the liver and spleen for phagocytic disposal (Mak et al., 2014). In addition to surface binding, C3b activates the terminal pathway of complement by binding to a C3 convertase. This generates a C5 convertase in the form of either a C4bC3b2a or C3b₂Bb complex. By this, C5 convertases cleave the complement protein C5 which generates activated C5b and the anaphylatoxin fragment C5a. For C5b, reaction with the complement proteins C6, C7, C8 and C9 generates the membrane-attack complex (MAC) C5b6789_n. For the lysis of pathogens, the MAC forms pores in their membrane which disrupts the osmotic balance and leads to cell death. For the released anaphylatoxin polypeptides, such as C3a, C4a and C5a, their pro-inflammatory functions include the activation of granulocytes, mast cells and macrophages (Klos et al., 2009). They also promote muscle contraction and increasing vascular permeability.

1.3 The classical pathway

The activation of the classical pathway involves the recognition of IgG- and IgM-containing immune complexes. The classical pathway is also activated independently of antibodies by direct interaction with C-reactive protein, serum amyloid P and apoptotic cells (Nauta et al., 2002). For the activation of C3 by the classical pathway, each of the pro-enzymes C1, C2 and C4 are activated. For C1, its multimeric C1q_{r2s2} complex is formed of a 462 kDa six-domain C1q molecule and two of each of 83 kDa C1r and C1s molecules. In C1, the C1q subcomponent is responsible for both recognition and binding whereas both C1r and C1s are pro-enzyme serine proteases. For the activation of C1, C1q interacts with either immune complexes, Ig activators such as IgG- or IgM-containing aggregates or polyanions such as bacterial liposaccharides, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Upon C1 activation, both of the C1r and C1s pro-enzymes auto-activate. For the activation of C4, auto-activated C1s catalyses the proteolysis of the C4 α -chain. This generates the anaphylatoxin C4a and the active C4b fragment. For the activation of C2, C4b binds to the C2b domain of C2 in the presence of C1s which splits C2 into the fragments C2a and C2b. Next, C4b binds to C2a which forms the classical pathway C3 convertase C4b2a complex capable of activating C3. For C4b, in addition to forming part of the classical pathway C3 convertase, its exposed thioester allows it to bind to cell surfaces by either amide or ester bonds (Campbell et al., 1980; Law et al., 1980) (Figure 1.2). This directs complement activation, via the formation of C4b2a complexes, to the target-cell surface. In addition, the opsonisation of target cells by C4b triggers phagocytic immune clearance via CR1 on phagocytes (Law & Reid, 1995).

1.4 The lectin pathway

The lectin pathway is activated by the detection of pattern recognition molecules such as complex carbohydrates on the cellular surfaces of bacteria and viruses by serum MBL. MBL consists of both carbohydrate binding modules and C1q-like triple-helical collagen-like regions. In plasma, MBL is complexed with MBL-associated serine protease (MASP). For MASP, it was recently shown that the MASP-1 form activates the MASP-2 form, and thus suggested that both MASP-1 and MASP-2 are essential for complement activation (Garred et al., 2016). The classical pathway C1q_{r2s2} and lectin pathway MBL-MASP complexes are structurally and functionally similar to each other.

Upon activation, MASP cleaves C4 and C2, which leads to the generation of the C3 and C5 convertases C4b2a and C4b3b2a.

1.5 The alternative pathway

In contrast to the classical and lectin pathways, the AP is spontaneously activated at a low rate by the water-mediated hydrolysis of C3. In this way, the AP provides an important mechanism of keeping complement alert which allows the constant probing of cell surfaces. This is termed “tick-over” (Bexborn et al., 2008). Water molecules are both small and nucleophilic enough to access and subsequently react with the buried thioester in C3. The hydrolysed C3(H₂O) (also known as C3i or C3u) complex is generated at a rate of 0.2-0.4% per hour (Pangburn & Muller-Eberhard, 1983), without the loss of the C3a fragment. C3(H₂O) is also short-lived (Law & Reid, 1995). The serine proteases factor B (FB) and factor D (FD) and the protein properdin are all activators of the AP. They are not part of the classical and lectin pathways. In the AP, FB circulates as a zymogen until it reacts with C3(H₂O) and forms a C3(H₂O)B complex. Cleavage by FD then activates the FB zymogen which yields the active but unstable C3 convertase C3(H₂O)Bb and the fragment Ba. The active Bb fragment (residues 235-739) is also a serine protease that consists of both a von Willebrand Factor type A domain and a serine protease domain. The Ba fragment (residues 1-234) (Fishelson et al., 1984) consists of three short complement regulator (SCR) domains and a linker. For FB to bind C3b or C3(H₂O), each of the Ba fragment, the Bb Mg²⁺-dependent metal ion-dependent adhesion site motif and Mg²⁺ ions are required. The C3(H₂O)Bb convertase can activate a few molecules of C3 to C3b before the complex rapidly decays (Lindorfer et al., 2010). In turn, this can initiate the formation of the C3bBb convertase (Figure 1.1). In addition to its hydrolysis by water, C3 in the AP is also activated slowly and continuously to C3b by either serum proteases or non-specific perturbations. These lead to conformational changes in C3 that expose the thioester. As with C3(H₂O), active C3b reacts with FB in the presence of Mg²⁺ ions to form C3bB, and FD reacts with C3bB to form the AP C3 convertase C3bBb and the Ba fragment. Both of the C3 convertases C3(H₂O)Bb and C3bBb have a short half-life of 90 s (Fishelson et al., 1984), and once Bb dissociates from C3b, it cannot re-associate (Pangburn & Muller-Eberhard, 1986). However, for the C3 and C5 convertases of the AP, their binding to properdin prevents both spontaneous dissociation and regulator-induced decay. Thus, properdin increases the stability of the convertases ten-fold on target surfaces (Fearon & Austen, 1975) and also inhibits its

regulation (Medicus et al., 1976). Properdin is the only known naturally occurring positive regulator of complement. Its importance in host defence is shown by that approximately 50% of properdin-deficient individuals experience severe bacterial infections that have a fatality rate of almost 75% (Schwaebble & Reid, 1999). If a target surface is in close proximity, C3b generated by either a C3(H₂O)Bb or C3bBb convertase can covalently bind to the surface and initiate AP activation. However, the probability of this occurring is very low because C3(H₂O) in the fluid phase is short-lived and both C3(H₂O) and C3b are regulated. These restrict both the production of C3 convertase and the deposition of C3b on host cells (Law & Reid, 1995). If a C3b molecule is deposited on the target surface, an “explosive” amplification loop of C3 activation occurs. Thus, on the target surface, C3b can react with FB in the presence of FD and properdin to generate a C3bBb C3 convertase which in turn activates more C3 to C3b. In the absence of any control, this leads to a positive feedback mechanism that depletes plasma C3 and leads to attack of the target. Interestingly, when compared to C3(H₂O)Bb, the convertase activity of C3bBb approximately doubled but both the stability and the resistance to inactivation by regulation decreased (Bexborn et al., 2008). In the AP, for C3(H₂O), its noncovalent binding to complement receptor type 2 promotes the formation of a membrane-associated C3 convertase C3(H₂O)Bb on itself. This focusses complement AP activation on complement receptor type 2-expressing cells, such as immune and epithelial cells, which is not completely understood in physiological terms (Schwendinger et al., 1997).

1.6 Regulation of complement

The complement system is regulated in order to prevent unwanted complement activation and inflammation on host cell surfaces. Thus, a fine balance between activating and regulating components ensures that complement activation is firstly restricted to pathogenic surfaces only and then turned off once the pathogen has been destroyed. For the regulation of the three activated complement pathways, C3b in both the fluid phase and at host cellular surfaces is subject to irreversible proteolysis by the serine protease factor I (FI) and a cofactor. This stops the interaction of C3b with effector immune components that can lead to target destruction. For C3b, FI splits the α -chain of C3b at two sites on either side of the thioester (Figure 1.3). This yields a ~174 kDa inactive C3b (iC3b) and a ~2 kDa C3f fragment. iC3b consists of two polypeptides which are linked by two disulphide bonds and unable to bind FB (Lambris et al., 1996) in order to form an AP C3 convertase (Pangburn et al., 1977). Despite being a direct consequence of C3b

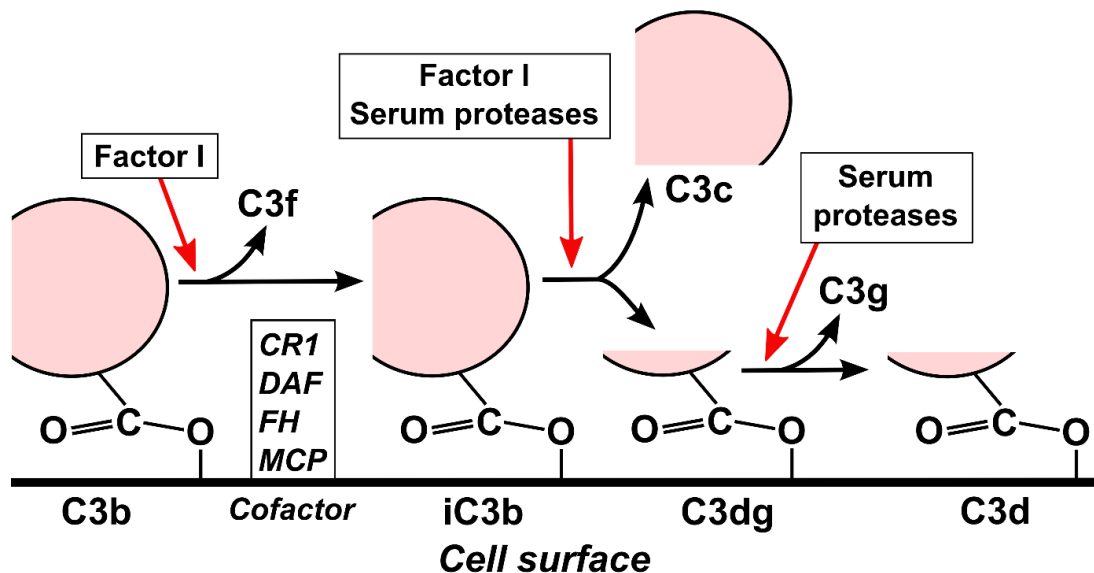


Figure 1.3 The inactivation of C3b. Surface-bound C3b on host cells is cleaved by factor I (FI) in the presence of one of the cofactors complement receptor 1 (CR1), decay-accelerating factor (DAF), factor H (FH) or (MCP). The first cleavage of C3b, at two places on the α chain (Arg1303-Ser1304 and Arg1320-Ser1321 ([Ross et al., 1983](#))), generates inactive C3b (iC3b) and C3f (residues 1304-1320). This disrupts the factor B binding site required for C3 convertase formation. iC3b consists of an intact β chain and an α chain which is in two fragments of 60 kDa and 40 kDa. Inter-chain disulphide bonds hold the α and β chains together. Further cleavage of the 60 kDa fragment of iC3b (at Arg954-Glu955) by either serum proteases or FI in association with CR1 generates C3c (23 kDa) and C3dg (37 kDa). C3c contains three chains: an intact β chain and two α chain-derived fragments (of 23 kDa and 40 kDa). Further cleavage of C3dg by serum proteases generates C3d (32 kDa) and C3g (5 kDa). The C3d fragment harbours the thioester thus remains attached the surface. Adapted from ([Law & Reid, 1995](#)).

inactivation, C3f is a 3 kDa weak spasmogen (i.e. smooth muscle contractor) with similar functions to the C3a anaphylatoxin ([Ganu et al., 1989](#)). Next, iC3b undergoes proteolysis by FI and serum proteases which generates the ~33 kDa C3dg and ~142 kDa C3c fragments. C3dg undergoes further proteolysis by serum proteases which generates C3d and C3g. At the host cell surface, C3d remains attached. A number of high-affinity complement receptors bind C3 fragments and generate either a pro-inflammatory response or tolerogenic suppression ([Merle et al., 2015a](#)), depending on the type of C3 fragment present. For example, the membrane-bound inactivation fragments iC3b, C3dg and C3d are ligands for complement receptor type 2 expressed on B cells, epithelial cells, follicular dendritic cells, thymocytes and some T cells. The interaction between these C3 fragments and complement receptor type 2 links the innate and the adaptive immune systems together and enhances the production of B cell-mediated antibodies ([Holers, 2014](#)). iC3b is also bound by complement receptors 3 and 4 expressed on macrophages, phagocytes, leukocytes and follicular dendritic cells. In addition, the C3 fragments are bound by an Ig superfamily complement receptor expressed on liver-resident macrophages which leads to the critical clearance of complement-opsonised fragments ([Helmy et al., 2006](#)).

FI and its cofactors are crucial for the regulation of the three activated complement pathways. FI is encoded by the *CFI* gene which has a genomic location of 4q25 ([Shiang et al., 1989](#)). *CFI* consists of 13 exons ([Figure 1.4](#)) and a 35 kb intron between exons 1 and 2 ([Vyse et al., 1994](#)). For *CFI*, instead of a TATA box promoter, an initiator element TCAGCCA is responsible for promoter activity. Both the initiator element TCAGCCA and the twice-occurring CTGGAT motif play an important role in its constitutive expression. After expression by hepatocytes in the liver and other cell types, FI is secreted as a multi-domain 88 kDa protein which circulates in human blood at 35 µg/mL (0.39 µM) ([Nilsson et al., 2011](#)). FI consists of a heavy 50 kDa and a light 38 kDa chain ([Goldberger et al., 1984](#)) linked by a disulphide bridge ([Fearon, 1977](#)). The heavy chain consists of a FI membrane attack complex domain, a scavenger receptor Cys-rich or CD5-like domain, low-density lipoprotein receptor 1 and 2 domains (LDLr1 and 2) and a small region of unknown homology which is sometimes called the D-region. A signal peptide of 18 residues is cleaved after secretion. For the light chain of FI, the serine protease domain ([Chamberlain et al., 1998](#)) harbours the catalytic triad formed by the residues His362, Asp411 and Ser507 in the active site ([Figure 1.4](#)).

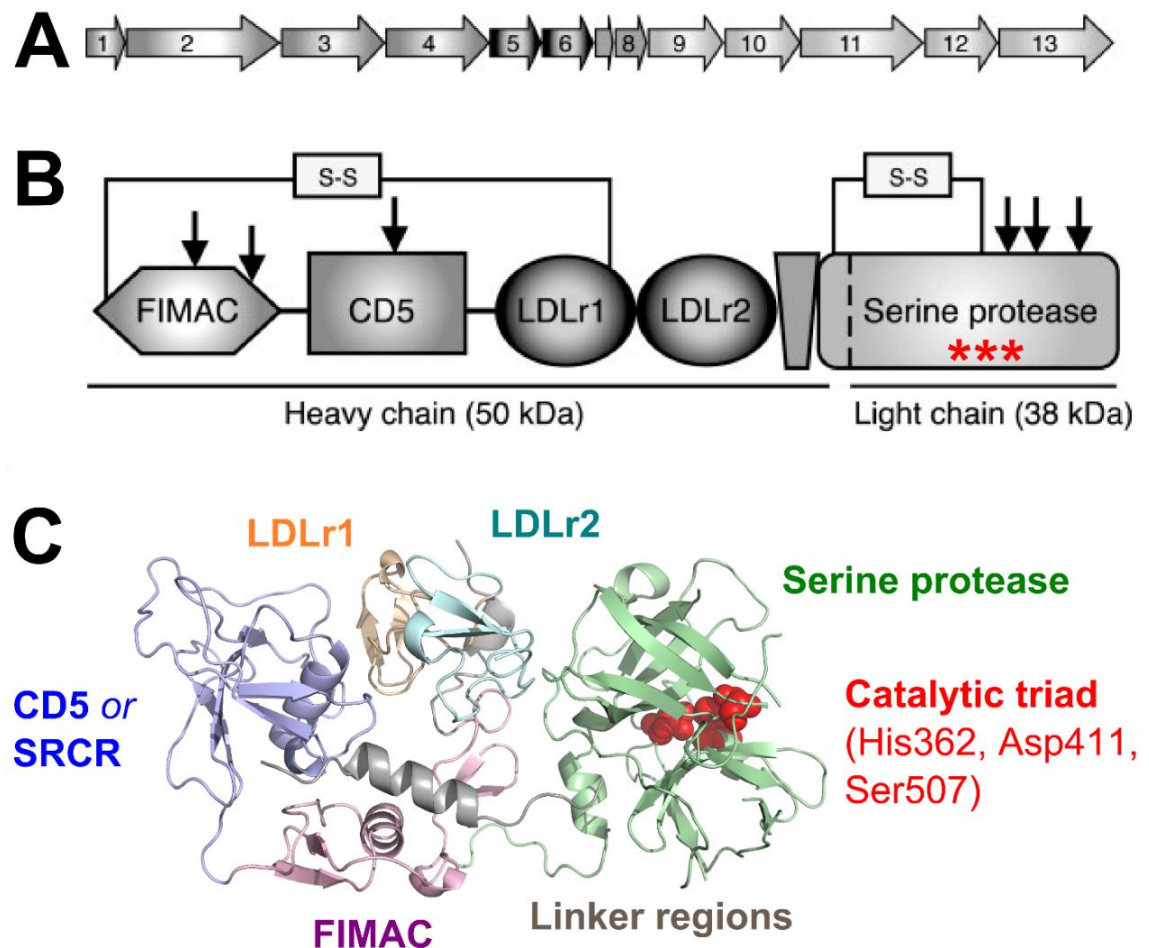


Figure 1.4 Structure of the complement factor I gene and protein. (A) For the complement factor I (*CFI*) gene, the arrangement of the 13 exons is shown. A very large intron of 35 kB is situated between exons 1 and 2 (not shown). (B) For the factor I (FI) protein, the five domains are schematically represented. The heavy chain consists of a FI membrane attack complex domain (FIMAC), a CD5-like or scavenger receptor Cys-rich domain and the low-density lipoprotein receptor 1 and 2 domains (LDLr1 and 2). The light chain consists of the serine protease domain which includes the active site formed by the catalytic triad. The arrows represent the locations of the six glycans on FI. (C) For the FI protein, the structural model is shown in cartoon-style and colour-coded by domain (Protein Data Bank code: 2XRC). The catalytic triad is shown as red spheres. Adapted from (Nilsson et al., 2011)

The cofactor used by FI depends on both the complement pathway and the local environment. For the FI-mediated proteolysis of C3b on host cell surfaces, the cofactors are membrane-bound CR1 (*CR1*), factor H (FH; *CFH*) and membrane cofactor protein (MCP; *CD46*). The *CR1*, *CFH* and *CD46* genes are all located within a cluster of 60 genes called the regulators of complement activation (RCA). The RCA cluster is located at the genomic location 1q32 and 15 of its 60 genes are complement-related. In the RCA cluster, *CR1* and *CD46* are in one group and *CFH* is in the other group, arranged in tandem (Rodriguez de Cordoba & Goicoechea de Jorge, 2008). The proteins encoded by RCA genes are not only related genetically but also structurally and functionally. Thus, the RCA family of proteins possess both SCR domains, which are ~60 residues in length and include four invariant Cys residues, and the ability to bind C3b or C4b. Overall, RCA proteins have major roles in the protection of host cells from complement by regulating complement activation, processing immune complexes, microbes and other foreign materials, and activating immune cells. This includes the binding, transport, and clearance of foreign material, endocytosis and signal transduction (Hourcade et al., 1989).

For the classical and the lectin pathways, the FI-mediated proteolysis of C4b relies upon cofactor activity from either C4 binding protein (C4bp; *C4BP*) (Fujita & Nussenzweig, 1979), CR1 or MCP. The *C4BP* gene is located within the first group of the RCA cluster (Rodriguez de Cordoba & Goicoechea de Jorge, 2008). The major isoform of C4bp consists of seven identical α -chains and one β -chain, which fold as eight and three SCR domains, respectively. C4bp mostly circulates with the anti-coagulant protein S which enables it to interact with negatively charged phospholipid membranes and dead cells (Webb et al., 2002). However, C4bp can also bind and protect host cell surfaces via heparan sulphate (HS) and its analogue heparin. C4bp binds to both C4b and heparin via its SCR-1/3 domains in the α -chain (Blom et al., 2001; Blom et al., 1999; Helsing et al., 1990). For C4bp SCR-1/3, heparin binding inhibited the binding of Ig-like surface proteins of the pathogenic bacteria *Leptospira* (Breda et al., 2015). In addition, C4bp bound to tumour cell surfaces via heparin was functionally active for FI-mediated cofactor activity (Holmberg et al., 2001). As for C3b, the α -chain of C4b is cleaved by FI at two sites on either side of the thioester, in the presence of a cofactor. This yields the C4c and C4d fragments. Currently, the functions of C4c and C4d are unknown and do not appear to include ligand binding. In addition to FI-mediated cofactor activity, both C4bp and membrane-bound CR1 can accelerate the decay of the classical C3 and C5 convertases (Ermer et al., 2013; Fujita & Nussenzweig, 1979). This is termed decay-

accelerating activity (DAA). Membrane-bound decay-accelerating factor (DAF) (Medof et al., 1984) also shows DAA for the classical C3 and C5 convertases. Thus, C4bp, CR1, and DAF bind to C4b and either block the formation of C4b2a and C4b3b2a or accelerate the dissociation of C2a from C4b (Law & Reid, 1995). C4bp can also act as a co-factor for the FI-mediated cleavage of fluid phase C3b, but cannot inhibit the AP C3 convertase (Seya et al., 1985) or efficiently regulate cell-bound C3b unless at a very high concentration (Blom et al., 2003; Fujita & Nussenzweig, 1979). Thus, for AP activity, C4bp has some control but cannot replace that of the AP regulator FH (Blom et al., 2004).

CR1 is a ~250 kDa membrane protein (Park et al., 2014) which regulates all three of the complement pathways. CR1 is composed of 30 SCR domains, a transmembrane and an intracytoplasmic domain. In addition to FI-mediated co-factor activity, CR1 can bind to either C3b or C4b thereby accelerating the dissociation of both the C3 and C5 convertases (Java et al., 2015). For CR1 on host immune cell surfaces, binding to C3b or C4b on opsonised pathogenic activator surfaces also triggers downstream immune processes. In contrast, for CR1 on host epithelial cell surfaces, binding to C3b or C4b locally deposited on the same surface allows the FI-mediated proteolysis of C3b for protection from complement attack. Indeed, for CR1 on host immune complexes, attachment to C3b can also degrade C3b to iC3b (Medof et al., 1982). On CR1, SCR-1/3 binds mainly C4b and has DAA, while SCR-8/10 and SCR-15/17 bind both C3b and C4b and have FI cofactor activity (Smith et al., 2002).

MCP is a membrane-associated protein which is encoded by the *CD46* gene. *CD46* has a genomic location of 1q32.2 and is part of the RCA cluster. *CD46* has 14 exons and is alternatively spliced to give rise to four major transcripts for MCP. By this, the alternatively spliced regions of *CD46* correspond to three exons which encode the O-glycosylation region and other exons which encode the carboxyl terminal cytoplasmic tail. For most cell types, the expression ratio of the four proteins is very similar, including on peripheral blood cells (Liszewski & Atkinson, 2015). MCP is composed of four SCR domains, the variable O-glycosylation region, a Ser/Thr/Pro-rich region, a transmembrane hydrophobic domain, a cytoplasmic anchor domain and the variable cytoplasmic tail domain. For all three of the complement pathways, MCP possesses cofactor activity for the FI-mediated proteolysis of C3b or C4b, but not DAA for the C3 or C5 convertases.

DAF is a ubiquitously expressed membrane-associated protein which is encoded by the *CD55* gene also located within the RCA cluster. Structurally, DAF consists of four SCR domains and a serine and threonine-rich domain which is bound to the membrane via a glycosylphosphatidylinositol anchor (Pan et al., 2015). For all three of the complement pathways, DAF possesses DAA for both the C3 and C5 convertases, but not cofactor activity for the FI-mediated proteolysis of C3b or C4b. DAF also plays a role in T cell and macrophage activation (Heeger et al., 2005).

Complement components other than C3b and C4b are also regulated. For the classical pathway, in order to prevent its activation in the absence of a C1 activator, C1 inhibitor (C1-Inh) rapidly forms a covalent complex with each of C1r and C1s (Sim & Reboul, 1981). By this, C1r and C1s are removed from C1q, which stops the activation of C4 by C1. C1-Inh also inhibits the proteolytic activities of MASP in the lectin pathway (Matsushita et al., 2000). For both the classical and the lectin pathways, the activator C2 is regulated at the cell membrane by C2 receptor inhibitor trispanning, which blocked C3 convertase formation *in vitro* (Inal et al., 2005). For all three pathways, after activation of C3 and C5, the complement anaphylatoxins C3a and C5a generated can both be inhibited by serum carboxypeptidases. By this, carboxypeptidase cleaves the C-terminal Arg residue, which generates C3a and C5a desArg, respectively (Matthews et al., 2004). The formations of both C3a and C5a desArg show a rapid loss of peptide and receptor binding activity. This was importantly shown to downregulate inflammation in autoimmune arthritis (Song et al., 2011). All three of the complement pathways are also regulated by sushi domain-containing protein 4 which is mainly expressed in the brain. For the regulation of the terminal pathway, each of the proteins protectin, clusterin and vitronectin inhibit either the formation or the membrane insertion of the MAC (Meri et al., 1990).

For the AP, in order to prevent host cell attack via its spontaneous activation, regulation is essential. Thus, FI and the cofactor FH regulate C3b in the fluid phase; and FI, membrane-bound FH and the three membrane proteins MCP, CR1 and DAF regulate C3b at the host-cell surface (Figure 1.1). FH is the main inhibitor of the AP (Rodriguez de Cordoba et al., 2004). FH regulates the AP in three ways. Firstly, FH is a cofactor for the FI-mediated degradation of either fluid phase or membrane-bound C3b, and C3b(H₂O) in the fluid phase. Secondly, FH has DAA for both of the C3 convertases C3bBb and C3(H₂O)Bb and also the C5 convertase C3b₂Bb, which leads to their

dissociations. In DAA, FH rapidly dissociates the enzymatic Bb domain from the C3 convertase. The DAA of FH is more effective for C3bBb than for C3(H₂O)Bb. The affinity of FH for C3(H₂O) is several fold weaker than for C3b (Heurich et al., 2011). Thirdly, FH blocks the formation of the C3bBb and C3(H₂O)Bb convertases. FH provides targeted regulation of AP activation at host cell surfaces by recognising surface poly-anions such as sialic acid and glycosaminoglycans. The cleavage of C3b by FI and the cofactor FH is inhibited by the complement activator properdin. In addition to AP regulation, FH can directly compete with C1q for binding to anionic phospholipids and act as a down-regulator of classical pathway activation (Tan et al., 2010).

1.6.1 Complement factor H

The complement regulator FH is encoded by the *CFH* gene. The genomic location of *CFH* is 1q31.3 which is within group 2 of the RCA cluster. Thus, *CFH* is located next to five homologous *CFH-related* (*CFHR*) genes that are situated in tandem (Figure 1.5). The *CFHR1-5* genes are thought to have arisen from gene duplication events. Their homologous repeat regions are prone to non-allelic homologous recombination events which can lead to chromosomal deletions and hybrid gene products. *CFH* is expressed from exons 1-9 spliced with exons 11-23 to form a full-length 150 kDa protein consisting of 1,213 residues (Rodriguez de Cordoba et al., 2004). Full-length FH is composed of 20 SCR domains which are joined by 19 flexible linker peptides of 3 to 8 residues (Okemefuna et al., 2009). *CFH* is alternatively spliced to form a truncated 42 kDa form which consists of 431 residues. This truncated form of FH is known as FH-like 1 protein (FHL-1) (Ripoche et al., 1988) and consists of the first seven SCR domains of FH plus four extra C terminal residues (Hellwage et al., 1997). The products of the *CFHR1-5* genes, the FH-related 1-5 (FHR1-5) proteins, are composed of different types and amounts of SCR domains. For example, FHR2 and one isoform of FHR4 are the shortest with four SCR domains, whereas FHR5 is the longest with nine SCR domains. FHR1-2 and FHR5 circulate in the blood as dimers, via the two N-terminal SCR domains, whereas FHR3-4 do not (Skerka et al., 2013).

In plasma, FH occurs both in the fluid phase and attached to host cell surfaces. FH is present in the plasma at a concentration of 110-615 µg/mL (~ 3.2 µM). The N-terminal SCR-1/4 domains of FH are involved in both C3b and FI binding, which are essential for both its regulatory cofactor activity and DAA. FH attaches to host cells by binding to

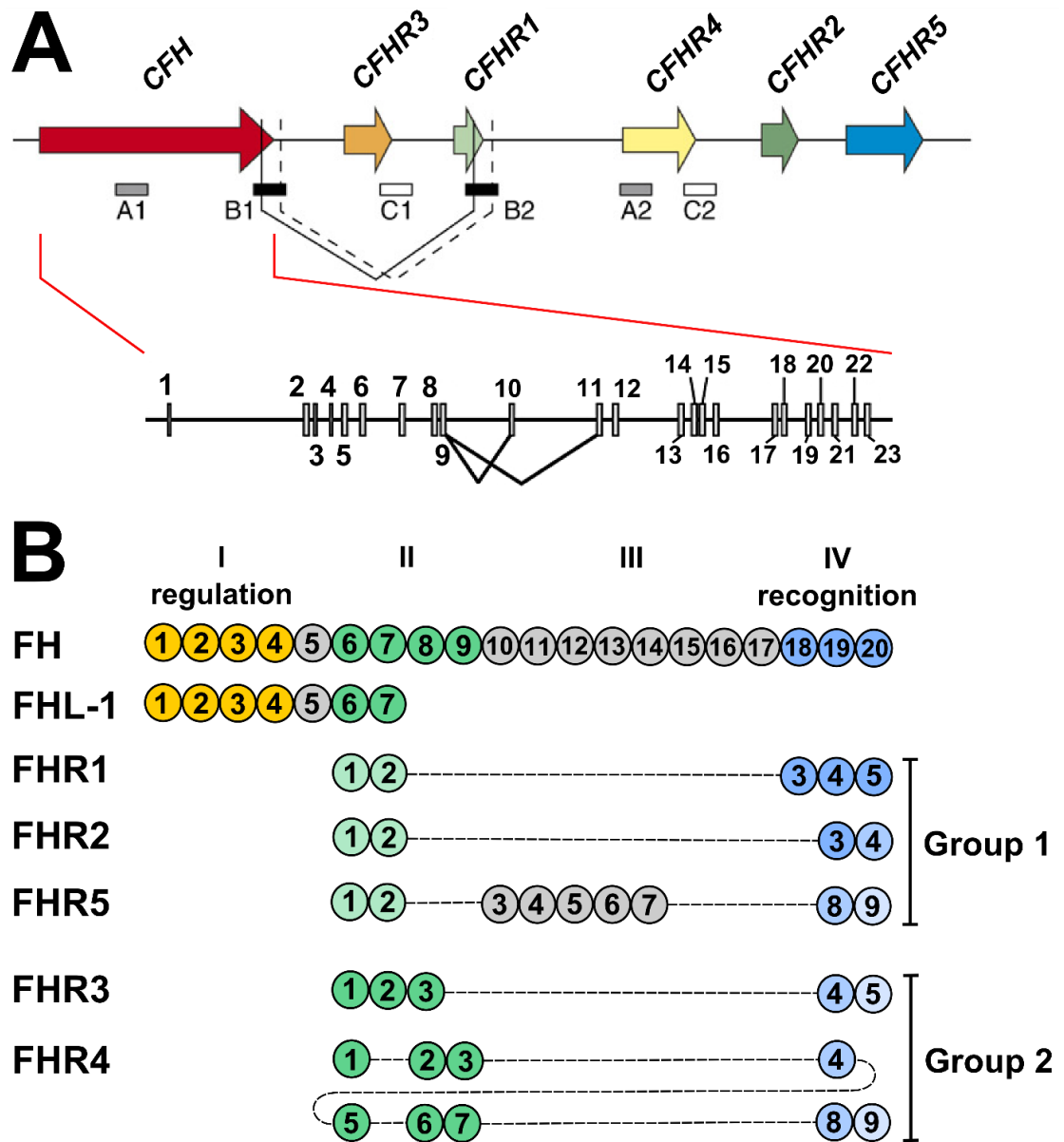


Figure 1.5 Schematic representation of the complement factor H gene cluster and protein family. Figure legend overleaf.

Figure 1.5 (continued) Schematic representation of the complement factor H gene cluster and protein family. (A) The human complement factor H (*CFH*) gene cluster on chromosome 1q32 includes the *CFH* and complement factor H-related (*CFHR*) 1-5 genes, in tandem. A1-C3 denote the homologous repeat regions which are involved in non-allelic homologous recombination events that can lead to large chromosomal rearrangements. For example, for B1 and B2, each of the solid and slashed black lines describes the *CFH*-*CFHR1* hybrid gene formation and the *CFHR1*-*CFHR3* gene deletion, respectively. The *CFH* gene consists of 23 exons which are alternatively spliced to generate either short complement repeat (SCR)-1/20 factor H (FH) (exons 1-9 and 11-23) or a SCR-1/7 FH-like (FHL-1) (exons 1-10) protein. (B) The FH protein family consists of FH, FHL-1 and five FHR proteins, which are all composed entirely of SCR domains. For FH and FHL-1 but not the five FHRs, SCR-1/4 in region I are associated with regulatory functions. For all six proteins, the N-terminal SCRs in region II (green) are similar in sequence. FHR5 SCR-3/7 have high sequence similarity to FH SCR-10/14 in region III. For FH and the five FHRs, the C-terminal SCRs in region IV (blue) share sequence homology and are associated with recognition functions. The FHRs in Group 1 have an N-terminal dimerization motif for the formation of homo-dimers whereas the FHRs in Group 2 do not. Adapted from (Rodriguez de Cordoba & Goicoechea de Jorge, 2008), (Jozsi & Zipfel, 2008) and (Skerka et al., 2013).

sialic acid and glycosaminoglycans such as HS on host cell surfaces. FH cannot bind to sialic acid on host cell surfaces in the absence of C3b. On FH, these interactions are mediated by the C-terminal SCR-18/20 domains and the middle SCR-6/8 domains. The C-terminal SCR-18/20 domains of FH are also involved in C3b and C3d recognition (Rodriguez de Cordoba et al., 2004).

In contrast, FHL-1 does not possess SCR-18/20 which are involved in host cell attachment and C3b and C3d recognition in FH. However, FHL-1 does possess SCR-1/4 and part of SCR-6/8 which are involved in regulatory activities and binding to host cell surfaces via HS, respectively. In terms of function, FHL-1 shows both cofactor activity for the FI-mediated degradation of C3b (Kuhn et al., 1995) and DAA for the AP C3 convertase (Kuhn & Zipfel, 1996). However, for DAA, FH is 100-fold more efficient than FHL-1. On the other hand, it has been suggested that the smaller size of FHL-1 may allow better tissue penetration which would affect local convertase regulation (Clark & Bishop, 2015).

For each of the five FHRs, despite possessing fewer SCR domains than FH, the SCR domains showed high sequence similarity (36–100%) to some of those in FH (Figure 1.5). For FH and the FHRs, the conservation of some of their SCR domain sequences suggests their related functions. For FHR1-5, the FH N-terminal SCR-1/4 domains are not conserved. Despite this, some of the FHRs still bind to C3b, but in a different way to FH. Thus, FHR1 cannot perform cofactor activity or C3 convertase (C3bBb) mediated DAA (Timmann et al., 1991), but can inhibit both the C5 convertase and formation of the terminal pathway complex (Heinen et al., 2009). In contrast FHR2 can inhibit both the AP amplification loop, by binding to the C3b in the C3 convertase via SCR-3/4, and the formation of the terminal pathway complex via SCR-1/2 (Eberhardt et al., 2011; Eberhardt et al., 2012). Recombinant FHR3 showed low cofactor activity for the FI-mediated cleavage of C3b which is probably not its main function. Both FHR3 and FHR4 may enhance the cofactor activity of FH (Hellwage et al., 1999). At high concentrations, FHR5 showed both cofactor activity and DAA. For all five FHRs, the N-terminal and C-terminal SCR domains showed high sequence similarity (36-100%) to the FH SCR-6/8 and SCR-18/20 domains, respectively. The FH SCR-6/8 and SCR-18/20 domains are involved in host cell recognition. Accordingly, the FHRs 1-5 can bind to C3b and C3d and discriminate between host and non-host surfaces (Skerka et al., 2013). For FHR5 only, five additional middle SCR domains show high sequence similarity (36-100%) to

SCR-10/14 of FH. For FH, the SCR-10/14 domains are involved in C-reactive protein interactions. For FHR5, C-reactive protein mediates its recruitment to damaged host cells in order to inactivate C3b ([Park & Wright, 1996](#)). Overall, the functions of the FHRs 1-5 encompass support for the cofactor activity of FH as well as additional complement regulatory activities such as inhibition of the terminal pathway. However, the FHRs can also compete with FH for C3b binding thus these will negatively impact FH function ([Skerka et al., 2013](#)).

1.7 Diseases of complement

Complement is a powerful surveillance system which rapidly detects and triggers the destruction of pathogens. However, complement can also be activated spontaneously by the AP, and activators themselves cannot discriminate between pathogens and the host. Thus, a fine balance between complement activators and regulators is essential to ensure that pathogens are killed whilst host cells are protected. If this balance is disrupted, complement will not be able to defend the host from either pathogens or itself, which can lead to either infectious or autoimmune diseases, respectively. Complement disease is classified as either primary (hereditary) or secondary (acquired) ([Grumach & Kirschfink, 2014](#)).

Secondary complement disease includes triggers of complement activation such as the massive influx of either pathogen-associated molecular patterns during sepsis or damage-associated molecular patterns during tissue injury. These amplify complement and can lead to complement activation on bystander host cells. This forms the basis of a condition known as systemic inflammatory response syndrome. Secondary complement disease may also be induced by biomaterials such as implants, accumulating cell debris from ageing and oxidative stress and autoantibodies. It may also involve protein aggregates that lead to the maintenance of complement-induced inflammatory states such as those described in models for Alzheimer disease, other neurological and neurodegenerative diseases and atherosclerosis ([Ricklin et al., 2016](#)). Secondary complement disease does not involve genetic variations in complement genes.

Primary deficiencies of or other altered functionalities within the complement system can disturb the balance of activators and regulators. These underlying genetic abnormalities in complement are often revealed after an immune trigger and the

manifestation of severe clinical symptoms. Primary deficiencies of complement represent ~1-6% of all primary immunodeficiencies (Ricklin et al., 2016). All complement deficiencies are rare, which is defined as affecting either less than one in 2000 (0.05%) or less than one in 1500 (~0.07%) people in the European Union and the United States, respectively (Boycott et al., 2013). Early complement component deficiencies are associated with autoimmune and immune complex-mediated diseases as well as recurrent bacterial infections. Thus, impaired clearance of immune complexes and apoptotic cells by the classical pathway may trigger the generation of autoantibodies. For example, deficiencies in the C1 complex, C2 and C4 of the classical pathway are all associated with systemic lupus erythematosus (Walport, 2002). Interestingly, for patients deficient in either C1, C4, or C2, a lower prevalence of infection was shown, when compared to C3 deficiency. This suggests that in the absence of classical pathway activation, an intact AP is able to activate some C3, which has been shown in patient serum experiments (Clark & Klebanoff, 1978; Frank & Sullivan, 2014). C3-deficient individuals have an increased susceptibility to infection by mainly encapsulated pyogenic bacteria such as meningococci, *haemophilus influenzae* and streptococci. For C3 deficiency, this can be caused by either reduced C3 expression or increased C3 consumption. The former can be caused by loss-of-function (LoF) mutation in C3 (Chapter 2, Section 2.4.3). The latter can be caused by the absence of C3 regulation, perhaps by either a gain-of-function (GoF) mutation in C3 or a LoF mutation in a C3 regulator. Thus, *in vitro* depletion of either FI or FH caused complete C3 consumption by the AP (Nicol & Lachmann, 1973). In patients with complete FI deficiency, exhaustive AP activation leads to both very low C3 and FB levels. FI deficiency is frequently associated with severe infections, glomerulonephritis and autoimmune diseases (Sadallah et al., 1999). For FH, previously known as β 1H, a haemolytic uraemic syndrome (HUS) patient and his unaffected brother showed a very low level of functional FH that caused low levels of both C3 and haemolytic complement via the AP. Their C4 levels were normal. It was suggested that a genetic defect in *CFH* was inherited in homozygosity from both of their heterozygous parents, who had half-normal *CFH* levels (Thompson & Winterborn, 1981). For the terminal complement components, C5 to C9, a deficiency increases the risk of meningococcal infection. In contrast, deficiencies in the classical pathway regulator C1-Inh form the basis of hereditary angioedema which is characterised by the accumulation of inflammation-related fluid in tissues.

In complement, most primary deficiencies are inherited as autosomal recessive traits, except for those in either C1-Inh or properdin which are autosomal dominant and X-linked, respectively (Frank & Sullivan, 2014). In humans, autosomal recessive traits require two copies of the affected gene whereas autosomal dominant traits require only one copy of the affected gene. Individuals with only one copy of an autosomal recessive trait are known as carriers. X-linked traits are carried by genes on the X chromosome only, thus males that inherit the X-linked trait are always affected. On the other hand, females that inherit the X-linked trait are only carriers. In summary, each of the activator and regulator components of complement are vital to the functioning of either all of complement (C3) or their respective pathways. Genetic deficiencies in any of these complement components can result in a wide variety of rare, severe diseases.

1.7.1 Atypical haemolytic uraemic syndrome

HUS is a thrombotic microangiopathy. In thrombotic microangiopathy, small blood vessels in mainly the glomeruli of the kidney (Figure 1.6) and the brain are occluded by thrombus (blood clots) resulting from endothelial injury. As for most thrombotic microangiopathies, HUS is characterised by a triad of haemolytic anaemia, thrombocytopenia and acute organ failure (Ruggenenti et al., 2001). The most common form of HUS is typical HUS which is caused by gastrointestinal infection by Shiga toxin-producing *Escherichia coli*. The Shiga toxin initiates the disease process by damaging endothelial cells (Jokiranta, 2017). Shiga toxin-producing *Escherichia coli*-HUS is thought to affect the kidneys the most due to the high expression of a receptor for the toxin on glomerular endothelium (Amaral et al., 2013). Low C3 levels are typical of HUS patients but not always seen.

Atypical HUS (aHUS) is not caused by Shiga toxin-producing *Escherichia coli* infection but by dysregulated complement activation. aHUS is an ultra-rare disease that is reported to occur at an incidence of approximately 0.5 per million per year (Goodship et al., 2017). In aHUS, an immune insult triggers complement activation, primarily by the AP, which cannot be properly controlled due to underlying genetic defects (Chapter 2, Section 2.11). This leads to endothelial cell attack and microvasculature occlusion (Figure 1.7) (Caprioli et al., 2006), which results in organ damage to the kidneys, gastrointestinal tract, liver, pancreas and brain. aHUS can occur at any age, and 10% of aHUS patients

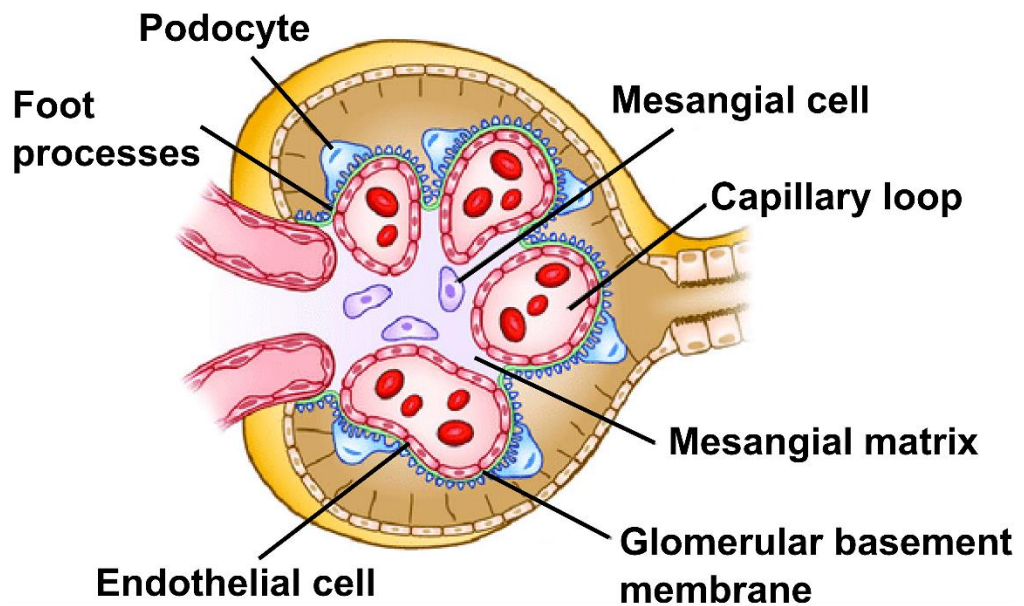


Figure 1.6 Morphology of the kidney glomerulus. The major functions of the kidneys include removing waste products and excess fluid from the body and regulating the salt, potassium and acid concentrations of the blood. The kidneys also release hormones that regulate blood pressure and control the production of red blood cells. Each of the two kidneys consist of millions of nephrons. A nephron consists of a tuft of capillaries, known as the glomerulus, surrounded by the Bowman's capsule. The glomerulus is a high-pressure capillary bed. To form urine, blood passes through the glomerulus and a filtrate collects in the Bowman's capsule via a filtration membrane. This filtration membrane is composed of the glomerular endothelial cells, the glomerular basement membrane, and the filtration slits between the podocytes. Sourced from ([Alicic et al., 2017](#)).

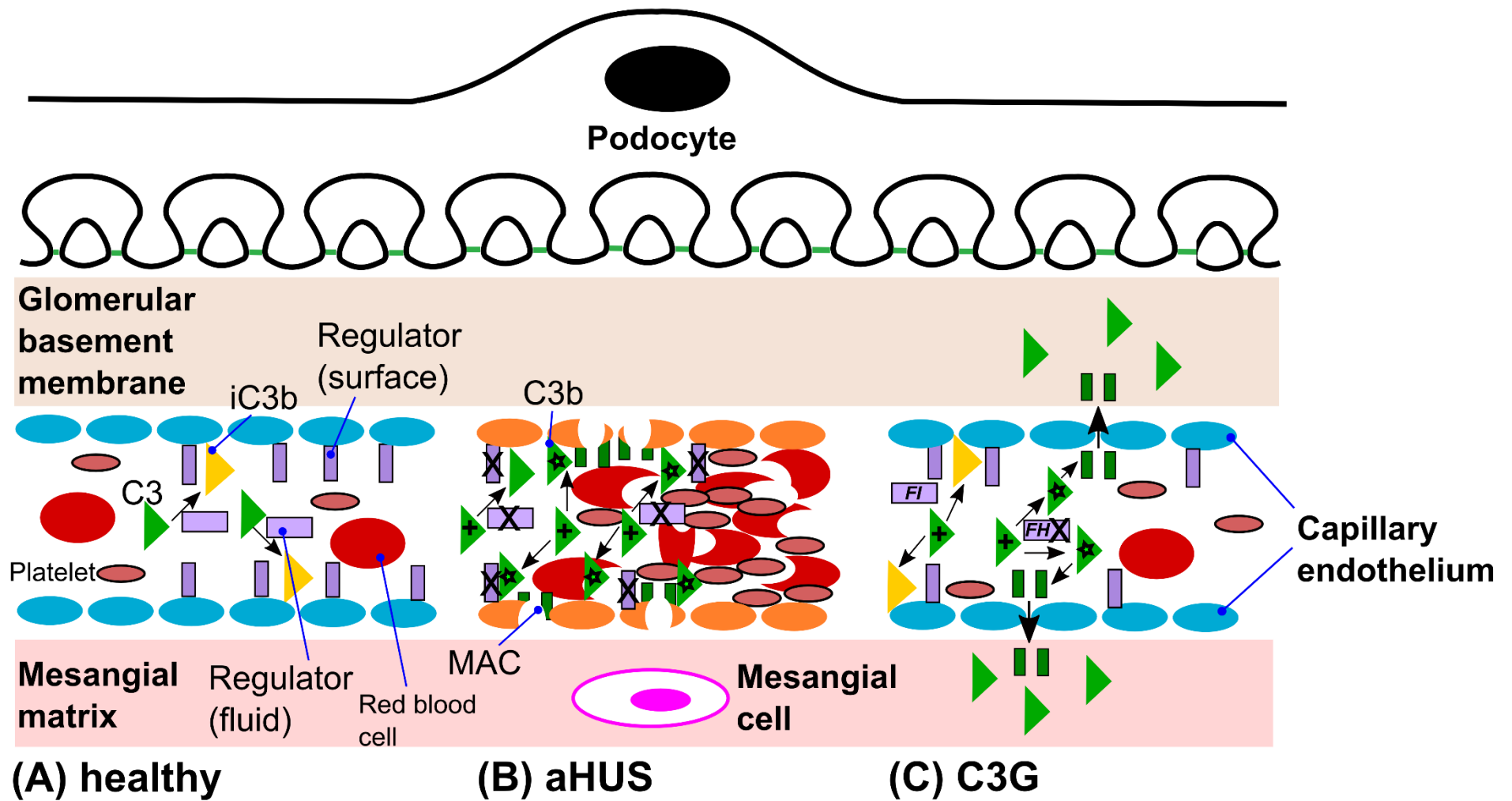


Figure 1.7 Models for the complement-based pathophysiology of aHUS and C3G in the kidney glomerulus. Figure legend overleaf.

Figure 1.7 (continued) Models for the complement-based pathophysiology of aHUS and C3G in the kidney glomerulus.

(A) In a healthy glomerulus, the low level tick-over activation of C3 (green triangle) to C3b by the complement alternative pathway (AP) is stopped by regulators (purple rectangles) both in the fluid phase (factor I (FI), factor H (FH)) and membrane-attached (membrane cofactor protein (MCP)) to the capillary endothelium. By this, host-cell deposited or fluid-phase C3b (green triangle) is inactivated to iC3b (yellow triangle) and host cells are protected from complement. Platelets activation (small, pink ovals), red blood cells (large, red ovals) and endothelial cells are normal and healthy (blue ovals).

(B) In atypical haemolytic uraemic syndrome (aHUS), dysregulation of the AP occurs due to loss-of-function (deficiency) in the regulators (FH, FI or MCP; purple rectangles with black cross), gain-of-function in the activators (C3, factor B; green triangles with black plus symbol) or acquired factors (auto-antibodies; not shown). This leads to impaired protection of host cell surfaces from AP activation of C3 to C3b (green triangles with black star), including membrane-attack complexes (MAC; dark green, double rectangles), inflammation and endothelial cell damage (orange ovals). By this, red blood cells are damaged (haemolytic anaemia; red, sickle-shaped ovals), platelets are activated (thrombocytopenia; small pink ovals) and the blood clots. For diacylglycerolkinase (DGKE)-associated aHUS, DGKE deficiency leads to a prothrombotic state (not shown).

(C) For C3 glomerulopathy (C3G), which can be further categorised into dense deposit disease (DDD) and C3 glomerulopathy (C3GN), electron-dense deposits of mostly C3 are seen in the glomerulus. These occur due to fluid phase AP dysregulation by either genetic or acquired factors. For C3G, the known predisposing genetic factors are low levels of FH or impaired FH cofactor activity (purple rectangle with black cross), and C3 gain-of-function (green triangle with black plus symbol), which lead to massive activation of C3 to C3b (green triangles with black star). This leads to deposition of C3 activating products and MAC in the subendothelial space upon transfer through the fenestrated endothelium ([Noris & Remuzzi, 2015](#)). For DDD, the deposits occur in the glomerular basement membrane, and may also occur in the mesangium (mesangial matrix and cells; both green triangles and dark green double rectangles). The accumulation of iC3b along the glomerular basement membrane (not shown) may be critical for the development of DDD ([Fakhouri et al., 2010](#)). For C3GN, only mesangial deposits are seen (green triangles and dark green double rectangles). For complement factor H-related 5 nephropathy, which may lead to a loss-of-function for FH, histologic appearances include mesangial and/or capillary wall C3 deposits ([Barbour et al., 2013](#)) (green triangles).

are children. Despite multiple genetic or acquired risk factors ([Esparza-Gordillo et al., 2006](#)), aHUS may not be manifested until the triggering stimulus occurs in middle age. For aHUS in childhood, the numbers of affected females and males are approximately equal ([Sellier-Leclerc et al., 2007](#)). However in adults, the literature suggests that more females than males are affected by aHUS ([Sullivan et al., 2010](#)). The occurrence of aHUS is either familial (<20%) or sporadic ([Bu et al., 2012](#)). Familial forms are defined as the presentation of aHUS in at least two members of the same family with individual diagnoses at least 6 months apart. For familial aHUS cases, the pattern of inheritance is either autosomal recessive or dominant ([Noris & Remuzzi, 2009](#)). The role of the AP in aHUS was first suggested in 1974 ([Kavanagh et al., 2008](#)). The first discoveries of the role of complement genetic variants in aHUS were in *CFH* in 1981 ([Thompson & Winterborn, 1981](#)) and 1998 ([Warwicker et al., 1998](#)). Presently, in ~60% of aHUS patients, one or more genetic abnormalities in complement AP or related genes have been detected ([Chapter 2, Section 2.11](#)). For acquired factors of aHUS, anti-FH autoantibodies are involved in 5–13% of aHUS cases. These are typically associated with homozygosity for a *CFHR3-CFHR1* genetic deletion.

Up until recently, the prognosis of aHUS was poor with patients progressing to end-stage renal disease within 2 years of diagnosis. Recently, the humanized monoclonal antibody Eculizumab (Soliris; Alexion Pharmaceuticals Inc), which inhibits C5, was shown to reverse the exacerbation of aHUS and control microangiopathic haemolytic activity ([Lapeyraque et al., 2011](#)). For aHUS, guidelines for the application of eculizumab were established in 2012 ([Zuber et al., 2012](#)) and eculizumab has since been associated with significant time-dependent improvement in renal function ([Legendre et al., 2013](#)).

Another form of HUS known as secondary HUS is caused by a co-existing disease or condition such as infection, transplantation, cancer, pregnancy, cytotoxic drugs or autoimmune diseases. These conditions can cause direct endothelial cell damage and promote complement activation ([Jokiranta, 2017](#)). Secondary HUS does not involve genetic abnormalities that affect complement regulation and is therefore a different condition to aHUS.

1.7.2 C3 glomerulopathy

C3 glomerulopathy (C3G) describes a spectrum of renal diseases that involve either genetic or acquired AP dysregulation, and leads to kidney failure. C3G is also ultra-rare with an incidence of approximately one per million per year (Goodship et al., 2017). However, unlike aHUS, C3G is not a thrombotic microangiopathy and C3G patients do not demonstrate anaemia, thrombocytopenia, or multisystem involvement (Bajwa et al., 2018). In C3G, uncontrolled complement activation leads to deposits of C3 and its fragments within the kidney glomerulus and its damage (Figure 1.7). Approximately 20% of C3G cases are associated with predisposing rare variants (RVs) in the complement genes (Chapter 2, Section 2.11). For C3G, diagnosis requires a renal biopsy and the C3 deposits are studied by electron microscopy (EM). Igs are absent from these deposits, which suggests that complement is activated independently of the classical pathway. Despite C3G and membranoproliferative glomerulonephritis (MPGN) sharing histopathological features, MPGN is caused by immune complex-mediated dysregulation of the classical pathway. C3G was classified in 2010 to reflect these differences (Fakhouri et al., 2010). C3G is further classified into either dense deposit disease if the C3 deposits are dense, osmiophilic and intramembraneous, or C3 glomerulonephritis if the deposits of C3 are a combination of light dense, mesangial, subendothelial or subepithelial (Figure 1.7). Most dense deposit disease (80-85%) and many C3 glomerulonephritis patients (~50%) develop acquired autoantibodies to the AP C3 convertases which are known as C3 nephritic factors (Servais et al., 2012). C3 nephritic factor protects the AP C3 convertase and leads to AP dysregulation (Appel et al., 2005). Anti-*CFH* autoantibodies are also acquired risk factors for C3G. An additional form of C3 glomerulonephritis, *CFHR5* nephropathy, is characterised by the presence of abnormal FHR5 (Chapter 2, Section 2.11). Currently, the prognosis of C3G is poor with most C3G patients progressing to end stage renal disease within a decade. For renal allografts, the recurrence rate of C3G is 45-60%. For C3G patients, the response rate to eculizumab is heterogeneous and not as successful as for aHUS (Welte et al., 2018).

Studies in mice have shown that the targeted deletion of *CFH* results in uncontrolled AP activation, very low circulating levels of C3, deposition of C3 on the glomerular basement membrane and MPGN-like characteristics of the glomerulus (Pickering et al., 2002). Further studies showed that mice with a transgenic *CFH* without the five C-terminal SCR domains were able to regulate the AP in the circulation but not

on the endothelial surface (Pickering et al., 2007). This lead to a renal thrombotic microangiopathy similar to aHUS (Cook, 2017).

1.7.3 Age-related macular degeneration

Age-related macular degeneration (AMD) is a common eye disease which affects ~50 million people across the world and is the leading cause of blindness in developed countries (Clark et al., 2014). For AMD patients, the loss of central vision can have a significant impact on their quality of life. Age is a major risk factor for AMD, alongside genetics and environmental factors such as nutrition. In AMD, with age, more polymorphous debris known as drusen are deposited between the retinal pigment epithelium (RPE) and Bruch's membrane (Figure 1.8). Also with age, the capability of the RPE to phagocytise debris likely decreases which adds to the accumulation of drusen. Under the RPE, drusen build-up is thought to activate complement and lead to inflammation and RPE damage. Drusen are classified by their size into small ($<63\ \mu\text{m}$), medium or large ($>124\ \mu\text{m}$) (Jager et al., 2008), which informs the grade of AMD. For example, medium-sized drusen in the absence of pigment abnormality are characteristic of early AMD for which there is a mild loss of vision. The molecular constituents of drusen include complement activators, components and regulators such as FH (Anderson et al., 2010). For the deposition of complement in drusen, photo-oxidised visual cycle products can activate complement. It has also been suggested that RPE cells might be prone to complement-mediated lysis (Ma et al., 2010). After the appearance of drusen, the development of AMD is subdivided into neovascular (wet) and atrophic (dry) forms. Similar to the complement-related diseases aHUS and C3G, complement genetic variants are thought to increase the risk of AMD by affecting the functionalities or circulating levels of the complement system. For AMD, this is expected to cause a hyperactive complement activation profile which, when combined with other factors, leads to excessive complement activation (Hecker et al., 2010; Heurich et al., 2011).

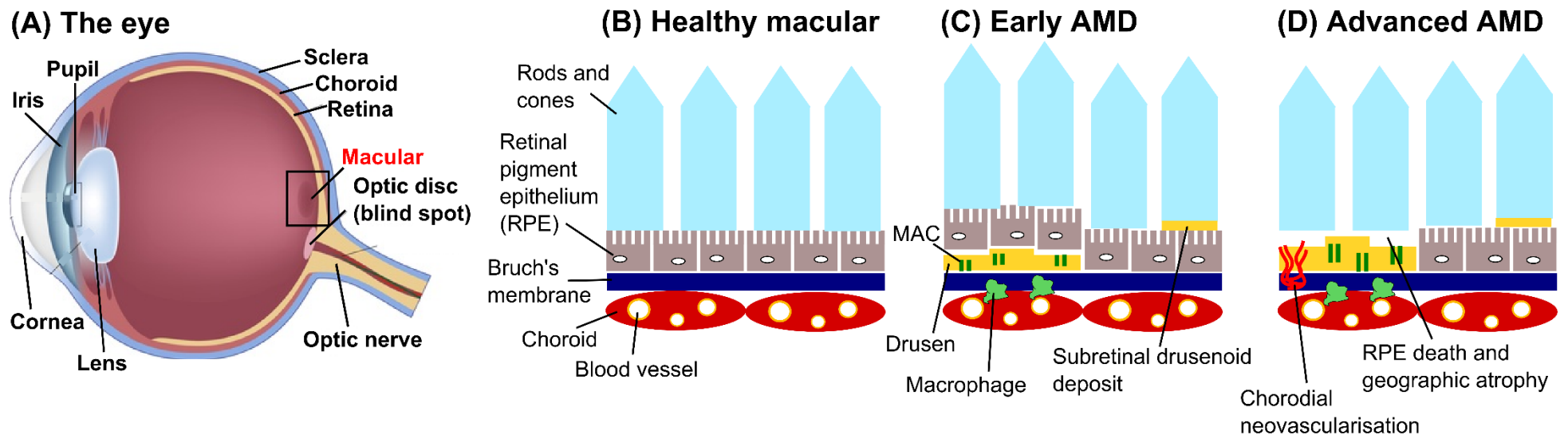


Figure 1.8 Age-related macular degeneration. (A) In the human eye, the macular is affected by age-related macular degeneration (AMD). (B) In a healthy macular, no inflammation or deposits (drusen) are observed. (C) In early AMD, drusen form between Bruch's membrane and the retinal pigment epithelium (RPE). Drusen contain the complement membrane attack complex and other complement components such as factor H. Macrophages are also present due to inflammatory processes. Slowly over time, the build-up of drusen increases loss of central vision. (D) In advanced AMD, severe sight loss is most commonly caused by either drusen (dry AMD) and less commonly by a combination of RPE cell death, geographic atrophy and the growth of new, abnormal and fragile blood vessels (wet AMD). In wet AMD, these vessels leak blood and fluid which damages the macula and makes central vision appear blurry. Adapted from (Apte, 2016) and an illustration provided by the National Library of Medicine (US) webpage [<https://ghr.nlm.nih.gov/condition/age-related-macular-degeneration>], accessed on 4 July 2018.

Chapter Two

Genetic variation

This introduction chapter describes the theory of genetic variation in humans, including how genetic variants can lead to disease phenotypes and how they are detected in the clinic by genetic sequencing. It provides a literature review of the role of complement genetic variants in the three diseases: atypical haemolytic uremic syndrome, C3 glomerulopathy and age-related macular degeneration. It describes bioinformatics methodologies for analysing the pathogenicity of these genetic variants, including allele frequencies, statistical analyses, web-databases and structural biology-based prediction tools. This chapter introduces my major research question for this PhD thesis, and outlines both the aims for my results [Chapters 4, 5 and 6](#), and methodology for [Chapters 4 and 6](#).

2.1 From DNA to proteins

One of the central goals of human biology and medicine is to understand the relationship between genotype and phenotype. For genetic diseases, the genotype-phenotype relationship allows underlying and/or predisposing factors to be identified. For patients with genetic diseases, the identification of genetic predisposition factors is highly useful and in some cases essential for predicting the outcome, prescribing treatment and the development of new therapies. For example, for aHUS, the identification of a pathogenic genetic variant reinforces the diagnosis and accurately establishes the cause of the disease. Overall, this facilitates patient management, effective treatment, such as anti-complement treatments, and genetic counselling. In addition, genetic analysis is essential for carrying out related-living donor organ transplantation such as for the kidneys in aHUS and C3G. Thus, the donor must not carry any genetic or acquired factors that have been clearly identified as causative in the recipient with the disease ([Goodship et al., 2017](#)).

In *homo sapiens*, or humans, the nucleus of every nucleated cell contains ~ 3 billion base pairs of the hereditary molecule DNA. This is known as the human genome. The first model of the double-helix structure of DNA was published in 1953 by Watson and Crick ([Watson & Crick, 1953](#)). For DNA, each nucleotide comprises one of four nucleobases which are classified by their structures into either purines (adenine, A and guanine, G) or pyrimidines (cytosine, C and thymine, T). The two separate strands of DNA are bound together by specific base pairing between A and T or C and G only.

According to the central dogma of biology, in a cell, DNA is transcribed into messenger ribonucleic acid (mRNA) which is then translated into one or a number of proteins ([Figure 2.1](#)). The translation of mRNA into a polypeptide of amino acids for protein folding is governed by a universal genetic code ([Table 2.1](#)). This genetic code shows degeneracy in that one amino acid can be encoded by more than one unique sequence of three DNA nucleotide bases. Overall, only ~1.5% of the genome corresponds to the ~20,000 genes that encode the human proteome. However, a further 80% of the genome is thought to be functional and may be transcribed, bind to regulatory proteins or be associated with other biochemical functions ([Pennisi, 2012](#)). In every cell, the proteome is different and constantly changes with time ([Munoz & Heck, 2014](#)). This is due to the differential expression of genes by different cell types. For the serum complement proteins, the main producers are hepatocytes in the liver. The only exceptions are FD, and factor properdin (FP) and C1q, which are primarily produced by adipocytes, and macrophages and other immune cells, respectively ([Morgan & Gasque, 1997](#)). Other cells such as endothelium and epithelium are also able to secrete various complement components, which may be essential for the localisation of complement at serum-restricted sites ([Lubbers et al., 2017](#)).

2.2 Genetic inheritance

The human genome is diploid and arranged into 23 pairs of chromosomes. Humans also have one copy of maternally inherited mitochondrial DNA. For the 22 non-sex chromosomes, also known as autosomes, two alleles or copies of the same gene are present at the same genomic position, or locus. Each one of these alleles is inherited from a parent. This mechanism of inheritance was first proposed by Gregor Mendel in 1865 in his principle of segregation, before DNA or genes were known to exist. Segregation refers to the separation of genes and their respective alleles during meiosis into daughter cells. Mendel also proposed the theory that one allele is dominant (A_1) whilst another is recessive (A_2).

For the phenotype, recessive traits require two copies (homozygous) of the allele whereas dominant traits only require one copy (heterozygous) of the allele. This applies to both the autosomes and the female sex chromosomes (XX) only. Dominant traits are observed in every generation whereas recessive traits are absent from one generation but

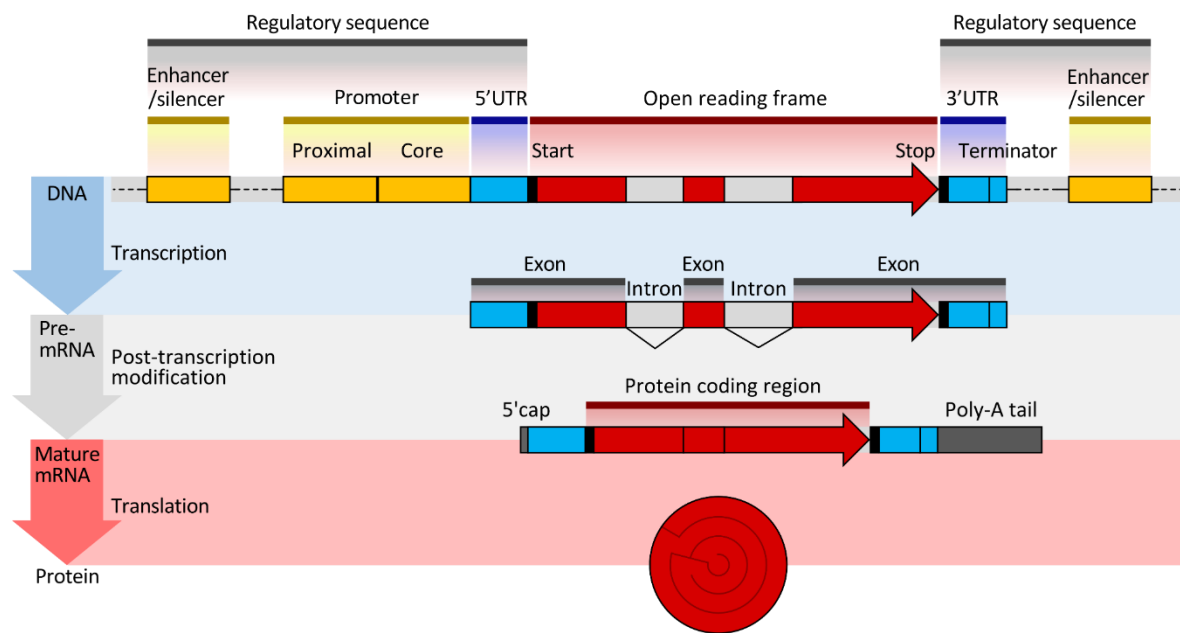


Figure 2.1 The central dogma of biology for a eukaryotic gene. Genes contain non-coding regulatory sequences such as the promoter (yellow) and the 5' and 3' untranslated regions (blue). Mature mRNA is capped at the 5' end and has a poly-adenine tail at the 3' end. Adapted from (Shafee & Lowe, 2017).

Table 2.1 The genetic code

Second base										
First base		T		C		A		G		
	T	TTT	Phenylalanine	TCT	Serine	TAT	Tyrosine	TGT	Cysteine	T
		TTC		TCC		TAC		TGC		C
		TTA	Leucine	TCA		TAA	Stop (<i>Ochre</i>)	TGA	Stop (<i>Opal</i>)	A
		TTG		TCG		TAG	Stop (<i>Amber</i>)	TGG	Tryptophan	G
	C	CTT		CCT	Proline	CAT	Histidine	CGT	Arginine	T
		CTC		CCC		CAC		CGC		C
		CTA		CCA		CAA	Glutamine	CGA		A
		CTG		CCG		CAG		CGG		G
	A	ATT	Isoleucine	ACT	Threonine	AAT	Asparagine	AGT	Serine	T
		ATC		ACC		AAC		AGC		C
		ATA		ACA		AAA	Lysine	AGA	Arginine	A
		ATG	Methionine	ACG		AAG		AGG		G
	G	GTT	Valine	GCT	Alanine	GAT	Aspartic acid	GGT	Glycine	T
		GTC		GCC		GAC		GGC		C
		GTA		GCA		GAA	Glutamic acid	GGA		A
		GTG		GCG		GAG		GGG		G

reappear in another. For heterozygotes, the phenotype can also be an intermediate state of the two heterozygous alleles, which is known as incomplete dominance, or an expression of both heterozygous traits in equal numbers, which is known as co-dominance. For the pair of sex chromosomes, X-linked recessive traits are seen in many more males than females, because both parents of the female must carry the X-linked trait, compared to only one required for males. For X-linked dominant traits, affected males can pass them to daughters only and not sons (Griffiths, 2000).

At any one locus, the composition of the two alleles is known as the genotype. At the population level, multiples alleles for one locus can exist. This forms the basis of genetic and phenotypic variation in a population. For two or more loci, the probability of their combined inheritance depends on the physical distance between them on one chromosome. For alleles at different loci, if the observed combination frequency is significantly greater than the expected frequency of random association, they are in linkage disequilibrium. By this, linkage disequilibrium refers to the non-random association of alleles at two or more loci (Robinson, 1998). For example, the human leukocyte antigen region displays strong linkage disequilibrium whereby certain human leukocyte antigen alleles are inherited together as a conserved haplotype. For *CFH*, haplotypes consisting of multiple polymorphisms are statistically associated with either susceptibility or protection for AMD (Li et al., 2006).

A single gene can also contribute to multiple phenotypic traits in a phenomenon known as pleiotropy. Non-genetic factors, termed as the environment, can also affect genetic variation and the phenotypic outcome.

2.3 Genetic variation and evolution

Genetic variation is defined as the difference in DNA sequences between individuals within a population. For humans, the average amount of genetic variation between any two individuals has been estimated at ~0.1 - 0.4% of the genome (Jorde & Wooding, 2004; Karki et al., 2015). This equates to 3 – 12 million base pairs, on average. Between different species, the amount of genetic variation is even greater. The ultimate source of all genetic variation is DNA mutation, which is essential for the process of natural selection (Barton, 2010) that can drive biological evolution and eventual speciation (Nei & Nozawa, 2011). The theory of biological evolution was published by

Charles Darwin in his book “On the Origin of Species” in 1859 ([Darwin, 1859](#)). For evolution, the five mechanisms are: mutation, natural selection, random genetic drift, gene flow and recombination. Most of the human genome has been shaped primarily by both mutation and random genetic drift ([Kimura, 1991](#)). Evolution is defined as heritable changes in the frequency of alleles in a population over time.

2.3.1 Allele frequency

For a variant of interest in a population, genetic sequencing and subsequent analyses ([Section 2.7](#)) allow the number of reference and alternative alleles to be determined. Following this, the frequency of each allele can be calculated by dividing the number of alleles of interest (allele count) by the total number of alleles in the population (allele number). For example, in aHUS, the complement factor H (*CFH*) missense variant c.157C>T; p.Arg53Cys was identified in heterozygosity in two patients (two T alleles) and in homozygosity in two patients (four T alleles), out of a total of 3128 screened patients (6256 T or C alleles). The resulting AF for this variant (T allele) is thus 0.096% ([Osborne et al., 2018a](#)). For each family, the allele is only counted once. Because of this, it is important to record any relations between patients.

Theoretically, without both evolutionary mechanisms such as new mutation and selection, and non-random mating, the allele and genotype frequencies of a population will remain constant between generations. This theory is described by the Hardy-Weinberg equilibrium (HWE). The HWE corresponds to a mathematical equation and was formulated by G. H. Hardy and Wilhelm Weinberg in 1908 ([Bacaër, 2011](#)). For the HWE, two different alleles at an autosomal locus that segregate in a population are denoted A_1 and A_2 . Their allele frequencies (AFs) are p and q , respectively:

$$f(A_1) + f(A_2) = p + q = 1 \quad (2.1)$$

The expected frequencies of the three possible genotypes for the homozygotes (A_1A_1 and A_2A_2) and heterozygotes (A_1A_2), respectively, are then:

$$f(A_1A_1) + f(A_2A_2) + f(A_1A_2) = p^2 + q^2 + 2pq = 1 \quad (2.2)$$

The total frequency of all alleles at a particular locus is equal to 1 or 100%. For a population, the HWE connects the frequency of an allele with the frequency of heterozygotes and homozygotes, as well as the genotype frequencies (Figure 2.2). For example, for the *CFH* single nucleotide polymorphism (SNP) rs800292 (c.184G>A; p.Val62Ile), for which the G allele was associated with AMD in an Iranian population, the observed genotype frequencies for the AMD group were 74% for GG homozygotes (A_1A_1), 22% for GA heterozygotes (A_1A_2) and 4% for AA homozygotes (A_2A_2) (Babanejad et al., 2016). The observed AFs for the G (p) and A (q) alleles are then:

$$p = (A_1A_1) + \frac{1}{2}(A_1A_2) = 0.74 + \frac{1}{2}(0.22) = 0.85 \quad (2.3)$$

$$q = 1 - p = 1 - 0.85 = 0.15 \quad (2.4)$$

From these AFs, in order to test for HWE, the expected genotype frequencies can be calculated. For example, the expected number of heterozygotes (A_1A_2) is:

$$(A_1A_2) = 2pq = 2(0.85)(0.15) = 0.255 = 26\% \quad (2.5)$$

Following this, the observed genotype frequencies are compared to the expected genotype frequencies by using a goodness-of-fit χ^2 test (Wittke-Thompson et al., 2005). If the difference is significant, the HWE is violated, and an evolutionary mechanism is inferred to have occurred in the population. For case-control studies (Section 2.8.2), in the case group, the HWE is often violated because a disease-associated allele is over-represented (Weiss & Scott, 2009). On the other hand, for the control or reference dataset, the HWE can typically be used as a quality control measure, whereby a major cause of HWE deviation is genotyping error among other causes. For other theoretical population studies, the HWE can also be used to detect cases of evolution, non-random mating and population stratification (Rousset & Raymond, 1995). For example, if the observed frequency of heterozygotes was significantly higher than expected, it may be associated with higher rates of survival than the homozygous genotype. In order to identify the evolutionary mechanism, further experiments can be planned.

For multiple alleles at any one locus in a population, such as for the three A, B and O alleles which determine blood group in humans, the HWE can also be applied (Kumar et al., 2014). For three alleles, there will be six genotypes. For one locus, the

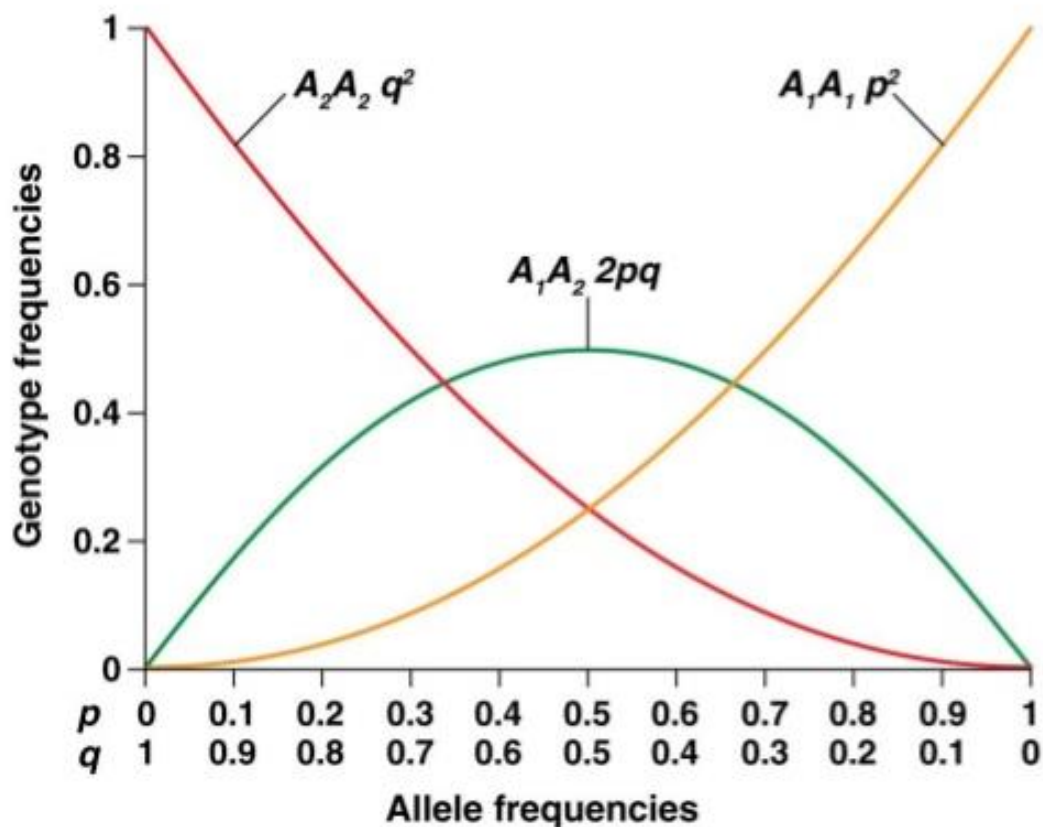


Figure 2.2 The Hardy-Weinberg equilibrium for two autosomal alleles. The relationship between genotype (A_1A_1 , A_1A_2 , A_2A_2) and allele (p , q) frequency. Heterozygotes (A_1A_2 , green) increase rapidly as the values of p and q move away from 0 or 1. For populations in which the frequencies of p and q are between 0.33 and 0.67, the frequency of heterozygotes (A_1A_2 , green) is higher than for either homozygote (A_1A_1 , yellow; A_2A_2 , red). The highest proportion of the population that can be heterozygous is 0.5 (green). Adapted from (Sanders & Bowman, 2012) and (Hedrick, 2005).

greater the number of alleles, the greater the heterozygosity, and the greater the genetic diversity (Rosenberg & Kang, 2015) (Figure 2.3). By this, the more variation that exists in a population, the better prepared the population will be for adaption to environmental change when it occurs.

2.3.2 Natural selection, genetic drift and gene flow

In natural selection, genotypes associated with high rates of survival and/or reproductive success are said to have high fitness, or genetic contribution, and are therefore selected for in the population. Positive selection maintains advantageous alleles whereas negative or purifying selection removes deleterious alleles. Selection can be weak or strong, depending on the magnitude of the advantage or disadvantage (Griffiths, 2000). However, deleterious alleles could be maintained at moderate to high frequencies in a population by balancing selection if they influence two or more independent phenotypic traits (pleiotropic) and at least one of these is a beneficial trait. By this, for a recessive allele that is associated with a genetic disease, the selection against homozygous recessives can be counter-balanced by selection in favour of heterozygotes (Altshuler et al., 2008). This is known as the heterozygote advantage, or over-dominance, and is seen for sickle-cell anaemia and perhaps cystic fibrosis. Here, the recessive allele also provides a survival advantage (Gemmell & Slate, 2006).

In genetic drift, the AFs of a population randomly change over generations in either direction. By this, a beneficial allele can become lost or a slightly harmful allele can become fixed (100%) in the population. These effects can significantly change the AFs and reduce genetic diversity. The smaller the population, the larger the effects of genetic drift. Two extreme examples of genetic drift are the bottleneck and the founder effect, both of which describe a severe reduction in the size of a population, due to either a catastrophe or isolation from the original population. Random genetic drift is thought to be responsible for most of the evolutionary changes at the molecular level, in the neutral theory of molecular evolution (Kimura, 1991).

Gene flow is described as the transfer of genetic variation from one population to another by inter-mixing of populations. Gene flow can also cause a change in AF.

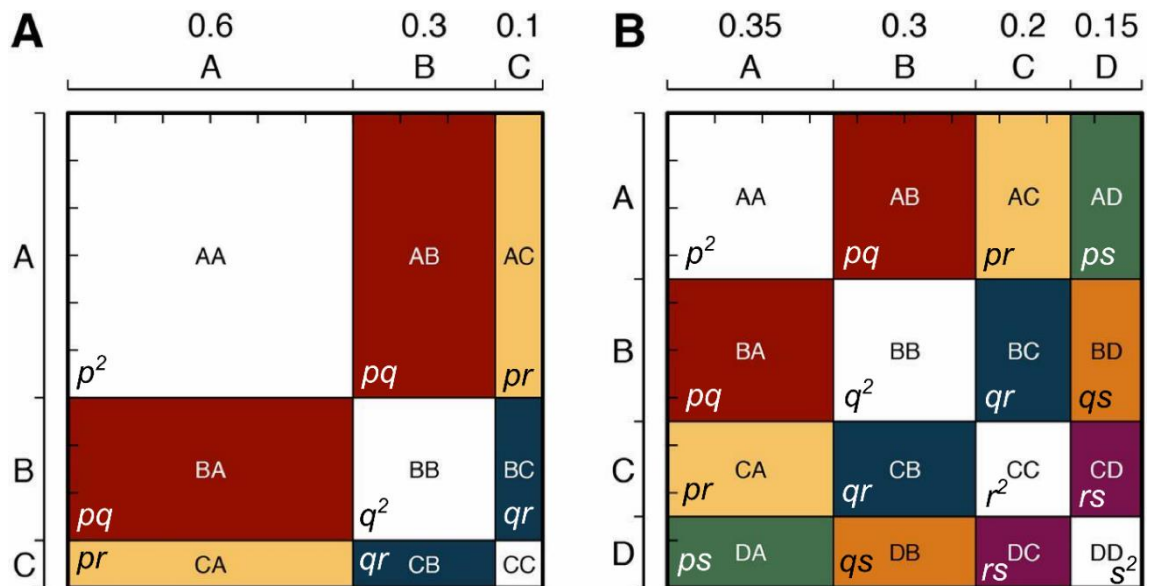


Figure 2.3 Punnett square showing the Hardy-Weinberg law for loci with three and four autosomal alleles in a population. (A) For one loci, the three alleles A, B and C are denoted p , q and r with example allele frequencies of 0.6, 0.3 and 0.1, respectively. For p^2 , q^2 , r^2 , pq , pr and qr , the genotype frequencies are thus 0.36, 0.90, 0.01, 0.18, 0.06 and 0.03, respectively. (B) For another loci, the four alleles A, B, C and D are denoted as p , q , r and s , at example allele frequencies of 0.35, 0.3, 0.2 and 0.15, respectively. For p^2 , q^2 , r^2 , s^2 , pq , pr , ps , qr , qs and rs , the genotype frequencies are thus 0.12, 0.90, 0.04, 0.02, 0.11, 0.07, 0.05, 0.02, 0.05, and 0.03, respectively. The white squares represent the homozygotes and the coloured squares represent the heterozygotes. Both the allele and the genotype frequencies add up to 1. Adapted from (Rosenberg & Kang, 2015).

2.3.3 Genetic recombination

Genetic recombination or exchange occurs via the transfer or copying of DNA segments during the pairing of the two homologous chromosomes in meiosis. Genetic recombination facilitates the formation of a new combination of alleles. For the adaptive immune system, genetic recombination drives rapid diversification of receptors and is essential for the recognition of new pathogens. In the genome, a lack of genetic recombination between sites strengthens the linkage disequilibrium.

2.3.4 Why do DNA mutations occur?

The ultimate source of all genetic variation is DNA mutation. Mutations occur due to either the spontaneous chemical instability of the bases, errors during DNA replication or environmentally by ionising radiation. Most mutations are believed to be caused by replication errors. During DNA replication, despite the high efficiency of DNA polymerase enzymes, incorrect nucleotides are added at a rate of approximately 1 per every 100,000 nucleotides. Thus, for each human diploid cell of 6 billion base pairs in cell division, approximately 120,000 nucleotide errors are made. Most of these single nucleotide errors are rapidly corrected by the proofreading or mismatch repair mechanisms during replication. However, some are not corrected and become permanent after the next cell division ([Pray, 2008](#)). For multicellular organisms, mutations can be inherited from parents (germline) or acquired over the life of an individual (somatic). The latter is often seen for cancer. Mutations in germinal tissue such as gametes may be transmitted to some or all progeny whereas those in somatic tissue are not transmitted to progeny. Within an organism, the presence of a genetically distinct (in terms of genotype) population of cells is known as mosaicism. Mosaicism can occur in either or both somatic and germline cells, depending on when the mutation occurred. There are six elementary types of DNA mutation: substitution, deletion, insertion, duplication, inversion and translocation. Each of these can affect the protein and its expression.

2.3.4.1 Substitution mutation

Substitution or point mutations in DNA involve the exchange of one single base pair for another. Substitutions are categorised into two types. Transition changes are when either a purine changes to the alternative purine (A to G, or G to A) or a pyrimidine

changes to the alternative pyrimidine (C to T, or T to C). Transversion changes describe either a purine to a pyrimidine (A to C, A to T, G to C, or G to T) or vice versa (C to A, C to G, T to A, T to G). During DNA replication, both transition and transversion mutations can be caused by tautomeric shift of the nucleotides which results in a mispairing. They can also be caused by bases which are either in different non-tautomeric chemical forms or slightly shifted in space, which also results in mismatched base pairing. This phenomenon is known as wobble and is facilitated by the flexibility in the DNA double helix. Wobble also provides an explanation for the degeneracy of the genetic code (Crick, 1966). In terms of spontaneous mutations, transitions can result from hydrolytic deamination of the pyrimidine C and also 5-methylcytosine involved in gene transcription, which generates uracil (U) and T, respectively. These deamination reactions have been shown to account for many single-base mutations that lead to human disease. For example, in the tumour suppressor gene *TP53* in carcinoma cells, C residues at CpG dinucleotides are mutational hotspots (Sassa et al., 2016). For the purines A and G, N-glycosidic bond hydrolysis leads to depurination which can also induce transition or transversion mutagenesis if not repaired by the base excision repair pathway. In humans, it has been estimated that around 2,000 – 10,000 DNA purine bases are turned over every day due to hydrolytic depurination and subsequent repair (Lindahl, 1993).

Theoretically, four transitions and eight transversions are possible. If the frequency of each nucleotide is the same, the expected transition to transversion ratio is 1/2. However, in humans, the transition to transversion ratio is approximately 2.0-2.1 for the whole genome and 3.0-3.3 for exonic variation (DePristo et al., 2011). Thus, transition substitutions are more favoured than transversions, which is known as transition bias. One reason for transition bias is that for the eleven amino acids that have two alternative codon forms (Phe, Tyr, His, Gln, Asp, Lys, Asp, Glu, Cys, Ser, Arg; Table 2.1), the alternate codons differ by one base at the third position and are either both purines or both pyrimidines (Higgs & Attwood, 2013). By this, transitions often result in synonymous (silent) changes which do not affect the fitness of the host and are therefore more likely to be tolerated. Furthermore, at the first and second codon positions, transition mutations tend to cause changes that conserve the chemical properties of amino acids (Wakeley, 1996). In addition, during replication, the generation of transversions requires greater distortion of the DNA helix than for transitions (Beamish, 2001). The transition to transversion ratio is a useful quality control parameter for genetic datasets.

For non-synonymous substitution mutations in the coding regions of DNA, the effect at the protein level is an amino acid substitution (missense mutation). A non-synonymous substitution mutation can also lead to the gaining of a premature stop codon (nonsense mutation), the loss of a stop codon, or the loss of an initiator codon (Table 2.2). Non-synonymous substitution mutations can also occur in the non-coding regions of DNA, such as the 3' and 5' untranslated regions (3'UTR and 5'UTR), introns, and non-coding transcript exons (Table 2.2).

SNPs are a type of substitution mutation which occur at an AF of more than 1% in a population. In humans, approximately 90% of sequence variation among individuals is due to common variants which includes SNPs (Kruglyak & Nickerson, 2001). SNPs are the most abundant form of genetic variation in the human genome and usually have a minimal impact on biological systems. However, SNPs can cause functional consequences such as changes to amino acids, mRNA transcript stability and transcription factor binding affinity. By this, SNPs can be risk factors for disease (Tabor et al., 2002). The frequency of a SNP is given in terms of the minor allele frequency (MAF), or the frequency of the second most frequent allele (Bush & Moore, 2012). For example, for the *CFH* SNP c.1204T>C which encodes the change p.Tyr402His, the MAF applies to the C allele and is reported to be 33% and 27% in each of the Exome Aggregation Consortium (ExAC) (Lek et al., 2016) and 1000 Genomes Project (1000GP) (Genomes Project et al., 2015) databases, respectively.

2.3.4.2 Other types of mutation

A deletion mutation can occur during DNA replication when the template strand slightly loops out and causes the deletion of one or several nucleotide bases from the newly synthesised DNA strand. This is known as strand slippage and is most often seen in regions of DNA with many copies of small repeated sequences (Pray, 2008). For a protein, a deletion of one or a few nucleotides could translate into either an in-frame or a frameshift mutation (Table 2.2), depending on the number of nucleotides involved. An in-frame mutation only results from nucleotides being in multiples of three. Otherwise, a frameshift mutation results.

An insertion mutation describes the addition of one or more nucleotides. In DNA replication, insertion mutations can be caused by the template strand looping out which

Table 2.2 Non-structural variant annotation

Annotation	Description	Example (for multiple genes)
3' UTR	3' untranslated region.	c.*10C>T
5' UTR	5' untranslated region.	c.-36G>C
frameshift	Any number of bases which are not a multiple of three are deleted or inserted which disrupts the reading frame from that point.	c.3286_3287delTG; p.Trp1096AspfsTer20
inframe deletion	Any number of bases which are a multiple of three are deleted which deletes any number of residues.	c.2422_2427delATACAA; p.Ile808_Gln809del
inframe insertion	Any number of bases which are a multiple of three are inserted which inserts any number of residues.	c.196_207insCAG...AGG; p.Gln68_Arg69dup
initiator codon	At least one base in the initiator codon is changed.	c.1A>G; p.Met1?
intron	At least one base in an intron is changed.	c.1336+136T>G
missense	A substitution mutation.	c.2465A>T; p.Asn822Ile
non coding transcript exon	A sequence variant that changes non-coding exon sequence in a non-coding transcript.	n.2587G>C
splice acceptor	A splice variant that changes the 2 base region at the 3' end of an intron (LoF).	c.2237-2A>G
splice donor	A splice variant that changes the 2 base pair region at the 5' end of an intron (LoF).	c.1519+5_1519+8delGT...
splice region	A splice variant that changes the region in between the two 2 base pair regions at the 3' and 5' ends of an intron.	c.1699A>C
stop gained	A nonsense mutation (LoF).	c.3643C>T; p.Arg1215Ter
stop lost	At least one base in the terminator codon is changed which changes the terminator residue.	c.1156T>C; p.Ter386ArgextTer69
stop retained	At least one base in the terminator codon is changed but the terminator remains.	c.205G>A; p.Ter69Ter
synonymous	At least one base is changed which doesn't affect the sequence but may affect the expression.	c.903A>G; p.Ala301Ala

Data sourced from the Exome Aggregation Consortium (ExAC) database.

results in strand slippage. A duplication mutation is the abnormal copying of one or more nucleotides in the DNA strand. For a protein, a duplication or insertion could translate into either an in-frame or a frameshift mutation ([Table 2.2](#)).

A complex mutation denotes a combination of two or more of the six different mutation types. One type of complex mutation combines an insertion and a deletion and is known as an indel. Thus, one or more nucleotides are deleted followed by the insertion of one or more different nucleotides.

An inversion mutation is where a segment of two or more nucleotides is reversed end to end. The overall number of nucleotides is not changed. Two chromosomal breakpoints must occur for an inversion.

A translocation mutation is where a segment of one or more nucleotides are moved to another genomic location between non-homologous chromosomes.

Structural variation is defined as a region of DNA which is approximately 1 kb or greater in size ([Freeman et al., 2006](#)). Structural variations can include genomic insertions and deletions, which are also commonly referred to as copy number variants, and inversions and translocations. Thus, copy number variants are a type of mutation in which the gene is amplified or reduced through increased or decreased copy number of a particular locus-specific allele ([Iafrate et al., 2004](#); [Sebat et al., 2004](#)). In addition to these types, segmental duplications are regions of DNA that are >1 kb in size and present more than once in the genome with >90% sequence identity ([Sharp et al., 2005](#)) and sometimes overlap with copy number variants. For structural variations, obtaining breakpoint resolution from sequencing is a major difficulty, and the events are highly variable. Thus, the Database of Genomic Structural Variation (dbVar) ([Lappalainen et al., 2013](#)) provides a database of human genomic structural variation to aid such analyses ([Table 2.3](#)). In non-allelic homologous recombination, a recombination event occurs between two lengths of DNA that have high sequence similarity (95-97%) but are not alleles. These events can cause either large chromosomal deletions or hybrid gene products, which lead to deletions or hybrid versions of one or more proteins. Non-allelic homologous recombination often occurs between DNA sequences that have been duplicated throughout evolution. For example, the *CFH* and *CFHR* genomic region ([Chapter 1, Section 1.6.1](#)).

Table 2.3 Structural variant annotation in dbVar

Variant Call Type	Description	Variant Region Type	Sequence Ontology ID
complex	A structural sequence alteration or rearrangement encompassing one or more genome fragments.	complex	SO:0001784
copy number gain/loss	A sequence alteration whereby the copy number of a given region is greater/less than the reference sequence.	copy number variation	SO:0001742 SO:0001743
deletion	The point of excision of one or more contiguous nucleotides.	copy number variation	SO:0000159
duplication	(copy number gain) A sequence alteration whereby the copy number of a given region is greater than the reference sequence.	copy number variation	SO:0001742
insertion	The sequence of one or more nucleotides added between two adjacent nucleotides in the sequence.	insertion	SO:0000667
inter-chromosomal / intra-chromosomal breakpoint	A rearrangement breakpoint between two different chromosomes / within the same chromosome	translocation or complex chromosomal mutation	SO:0001873 SO:0001874
inversion	A continuous nucleotide sequence is inverted in the same position.	inversion	SO:1000036
mobile element insertion	A kind of insertion where the inserted sequence is a mobile element.	mobile element insertion	SO:0001837
novel sequence insertion	An insertion the sequence of which cannot be mapped to the reference genome.	novel sequence insertion	SO:0001838
sequence alteration	A sequence alteration is a sequence feature whose extent is the deviation from another sequence.	sequence alteration	SO:0001059
short tandem repeat variation	A kind of sequence variant whereby a tandem repeat is expanded or contracted with regard to a reference.	short tandem repeat variation	SO:0002096
tandem duplication	A duplication consisting of 2 identical adjacent regions.	tandem duplication	SO:1000173
translocation	A region of nucleotide sequence that has translocated to a new position.	translocation	SO:0000199

2.4 How mutations affect proteins

In order to fully understand how a mutation will affect the function of a protein, thus lead to an altered phenotype, its effects on the structure of the protein are typically studied ([Chapter 3](#)). For a protein, the amino acid sequence (primary structure) is considered to carry information about both the structure and the function of the protein. Here, a brief overview of protein structure and the importance of each residue in defining the protein structure is given. For the primary structure of a protein, peptide bonds between consecutive amino acids form a polypeptide chain. For the secondary structure of a protein, patterns of hydrogen bonds between the C=O and H-N groups along the peptide backbone lead to certain types of local fold. For secondary structure, the dictionary of Protein Secondary Structure (DSSP) detects eight types based on the hydrogen bonding patterns and the electrostatic model ([Joosten et al., 2011](#); [Kabsch & Sander, 1983](#)). The most common types of secondary structure are alpha helices, turns and β -sheets in either anti-parallel or parallel form. For the formation of these secondary structure hydrogen bonds, the distance between the C=O and H-N groups is determined by the interaction of residue side chains with both the chemical environment and the polypeptide chain. For soluble proteins, the ‘Hydrophobicity’ and ‘Hydropathy Index’ properties of residue side chains ([Table 2.4](#)) drive the residues into positions which are either buried within the protein (hydrophobic) or on the protein surface in contact with water (hydrophilic). Thus, for soluble proteins, the global folding process is driven by the hydrophobic effect of water in the environment. For a transmembrane helix, the hydrophobic residues are typically in contact with the surrounding hydrophobic lipid bilayer. For residues, this allows both the solvent accessibility and associated secondary structure to be derived, e.g. residues from β -sheets are the most inaccessible to solvent, whereas random coils and turns are the most accessible to solvent. The residues in α -helices are approximately 20% accessible ([Lins et al., 2003](#)). For a residue in a protein, the relative solvent accessibility measures the extent of burial or exposure of that residue in the 3-dimensional (3D) structure. It is defined as the solvent accessibility normalized by a suitable maximum value for that residue. For determining protein structure and function, the relative solvent accessibility of each residue is often used in the first predictive stages ([Tien et al., 2013](#)).

For the polypeptide chain, both the size and chemistry of the residue side chains restrict the phi (ϕ) and psi (ψ) dihedral angles around the peptide bond due to steric clash.

Table 2.4 The 20 amino acids and their properties

Symbol	Code	Name	Type	Cyclic	Mass [Dalton]	Size	Hydrophobicity	Charge	Polarity	Surface Residue		
										Area [Å ²]	Volume [Å ³]	Hydropathy Index
A	Ala	Alanine	Aliphatic	Acyclic	71.09	Small	Hydrophobic	Neutral	Nonpolar	115	88.6	1.8
R	Arg	Arginine	Basic	Acyclic	156.19	Large	Hydrophilic	Positive	Polar	225	173.4	-4.5
N	Asn	Asparagine	-	Acyclic	115.09	Medium	Hydrophilic	Neutral	Polar	150	111.1	-3.5
D	Asp	Aspartic Acid	Acidic	Acyclic	114.11	Medium	Hydrophilic	Negative	Polar	160	114.1	-3.5
C	Cys	Cysteine	-	Acyclic	103.15	Medium	Hydrophilic	Neutral	Polar	135	108.5	2.5
E	Glu	Glutamic Acid	Acidic	Acyclic	129.12	Large	Hydrophilic	Negative	Polar	190	138.4	-3.5
Q	Gln	Glutamine	-	Acyclic	128.14	Large	Hydrophilic	Neutral	Polar	180	143.8	-3.5
G	Gly	Glycine	Aliphatic	Acyclic	57.05	Small	Hydrophobic	Neutral	Polar	75	60.1	-0.4
H	His	Histidine	Aromatic/Basic	Cyclic	137.14	Large	Hydrophilic	Positive	Polar	195	153.2	-3.2
I	Ile	Isoleucine	Aliphatic	Acyclic	113.16	Large	Hydrophobic	Neutral	Nonpolar	175	166.7	4.5
L	Leu	Leucine	Aliphatic	Acyclic	113.16	Large	Hydrophobic	Neutral	Nonpolar	170	166.7	3.8
K	Lys	Lysine	Basic	Acyclic	128.17	Large	Hydrophilic	Positive	Polar	200	168.6	-3.9
M	Met	Methionine	-	Acyclic	131.19	Large	Hydrophobic	Neutral	Nonpolar	185	162.9	1.9
F	Phe	Phenylalanine	Aromatic	Cyclic	147.18	Large	Hydrophobic	Neutral	Nonpolar	210	189.9	2.8
P	Pro	Proline	-	Cyclic	97.12	Medium	Hydrophobic	Neutral	Nonpolar	145	112.7	-1.6
S	Ser	Serine	-	Acyclic	87.08	Small	Hydrophilic	Neutral	Polar	115	89	-0.8
T	Thr	Threonine	-	Acyclic	101.11	Medium	Hydrophilic	Neutral	Polar	140	116.1	-0.7
W	Trp	Tryptophan	Aromatic	Cyclic	186.12	Large	Hydrophobic	Neutral	Nonpolar	255	227.8	-0.9
Y	Tyr	Tyrosine	Aromatic	Cyclic	163.18	Large	Hydrophobic	Neutral	Polar	230	193.6	-1.3
V	Val	Valine	Aliphatic	Acyclic	99.14	Medium	Hydrophobic	Neutral	Nonpolar	155	140	4.2

The data in this table was sourced from the PhD thesis of ([Rallapalli, 2014](#)).

Thus, for each residue, only certain combinations of dihedral angles are permitted, and this influences the secondary structure of the protein. For example, large aromatic residues prefer β -strand conformations whereas the residues Met, Ala, Leu, Glu and Lys (“MALEK”) prefer helical conformations (Table 2.4) (Myers et al., 1997; Pace & Scholtz, 1998). For α -helices or β -sheets, proline residues located internally can break the secondary structures and lead to protein destabilization because the amide proton is absent (Choi & Mayo, 2006). For a protein, the Ramachandran plot maps the dihedral angles of each residue onto an expected, empirically determined distribution of dihedral angles in order to detect any strained residues that are not likely to occur in a protein structure. Half of the regions of the Ramachandran plot are sterically inaccessible by residues (Haimov & Srebnik, 2016). The Ramachandran plot is useful for checking the quality of the secondary structure of a protein structure (Ramachandran & Sasisekharan, 1968). For the tertiary structure of a protein, one or more secondary structure formations can form independent stable functional domains (Branden & Tooze, 1999). Interactions between side chains such as disulphide bonds stabilise tertiary structure. In addition, the side chains of residues which are polar (Table 2.4) are able to either donate or accept a hydrogen bond with either other side chain or main chain atoms (Baker & Hubbard, 1984). One type of tertiary structure includes the TIM (triosephosphate isomerase) barrel, which is a conserved fold that consists of eight α -helices and eight parallel β -strands in alternation along the protein backbone. Alternatively, some folds are all α -helical or all β -strands in their secondary structure. For the three enzymes triacylglycerol lipase, cholesterol esterase and serine carboxypeptidase, despite their different biochemical functions, each of their active-site catalytic triads are provided by an α/β hydrolase fold (Orengo et al., 2003). For FI in complement, its Trypsin-like serine protease functionality is provided by three catalytic triad residues (His362, Asp411, Ser507) in the serine protease domain (Chapter 1, Figure 1.4). These three residues are optimally arranged for reaction via a β -barrel fold which is composed of mainly β -strands (Dawson et al., 2017). For proteins, one particular structural fold may have many biochemical functions, and one particular function may have many structures. In summary, for a protein, the secondary and tertiary structures are not only essential for protein stabilisation but can also provide functional folds that mediate specific activities such as catalysis and ligand binding. Each residue of the polypeptide chain thus contributes to the overall protein structure and can dictate its function.

For a missense mutation in a coding region, the greater the difference between the properties of the wild-type and mutated residues ([Table 2.4](#)), the more likely the secondary and/or tertiary structure is altered or even destroyed. For a protein, such structural effects can lead to either degradation, aggregation or affect the ligand binding properties. Synonymous coding mutations can affect gene expression and alter the levels of the protein. For a mutation in a coding region which substitutes a non-stop codon to a stop codon (TAA, TAG or TGA; [Table 2.1](#)), a nonsense mutation results. For the polypeptide chain, nonsense mutations can occur at any point and lead to truncation. For a truncated transcript, the ability of the protein to fold and/or function may be compromised. This depends on the location of the nonsense mutation with respect to the structural and functional elements of the protein chain. With exceptions, the closer the nonsense mutation is towards the 3' end of the polypeptide (the C terminus), the more likely the effects on the protein will be minimal as less of the protein is affected.

For in-frame deletions or insertions, one or more residues in a row are deleted or inserted, respectively, in the polypeptide chain. Thus, after the deletion or insertion point, the rest of the residues in the polypeptide chain remain. On the other hand, for frameshift deletions or insertions, all of the residues after the deletion or insertion can be changed. This can lead to rapid degradation of the protein.

For disorders of plasma proteins, a low level of protein with intact function is known as Type 1. By this, the protein was either not secreted or rapidly degraded. Type 2 is defined when the protein level is normal but with less functional activity ([Rodriguez et al., 2014](#)). With the relevant clinical data, mutations can be associated with either Type 1 or 2.

2.4.1 Non-coding mutations

Non-coding mutations can either affect the expression of genes or cause aberrant splicing such as exon skipping or cryptic splice site utilisation. Splice sites are known as cryptic when they are either dormant or only used at low levels, unless activated by either mutation or nearby authentic splice sites ([Kapustin et al., 2011](#)). More than 90% of disease-associated SNPs are located in non-coding regions of the genome. These include promotor regions, enhancers and non-coding RNA genes ([Hrdlickova et al., 2014](#)). Such mutations located within either introns or splice sites can lead to abnormal transcripts

such as the retention of introns and additional residues, which can cause protein truncations. For example, the *CFH* splice donor site variant InterVening Sequence (intron) 6+1 G>A is located at the 5' end of the sixth intron, and retention of this intron results in the addition of 30 amino acids to the protein before the stop codon. Thus, a truncated protein of 20% of the wild-type results ([Wagner et al., 2016](#)).

2.4.2 Residue conservation

Functional DNA sequences, such as those encoding proteins, tend to be actively conserved or inherited throughout evolution ([Frazer et al., 2001](#)). For any sequence, these conserved regions or blocks of residues are known as motifs. Motifs are biologically significant in terms of structure and/or function ([Huang et al., 2013](#)). Between two sequences, high sequence similarity of more than 30% identity over the entire length implies homology. This is because the probability of such similar sequences occurring independently due to chance is low. For two genes that are homologous, their separation occurred due to either speciation (orthologous), or genetic duplication (paralogous). In general, for two proteins, as their sequence identity (homology) decreases, both the topological differences of the protein backbones and the relative positions of corresponding side chains diverge ([Hilbert et al., 1993](#)). However, for proteins, large proportions can be structurally aligned despite their low sequence identity (e.g. <20%) ([Chothia & Lesk, 1986](#)). In order to predict the functional significance of a residue, a multiple sequence alignment of the sequence of interest (coding nucleotides or protein) and its homologs shows how conserved the residue is throughout evolution. Thus, highly conserved positions tend to be intolerant to substitution, due to their functional importance, whereas those with a low degree of conservation tolerate most substitutions. For the multiple sequence alignment, the chemical similarity of residues can also be taken into account. The homologous proteins compared could be from different species or in the same species. The DNA and protein sequences can be sourced from publicly accessible databases and tools ([Section 2.7.1](#)).

2.4.3 Loss and gain of function mutations

Mutations that diminish the function of the protein are known as LoF mutations. On the other hand, mutations that enhance the function of the protein are known as GoF mutations. For these terms, the loss or gain of function refers to the protein and not always

its corresponding biological system. For example, a LoF mutation in the C3 protein may increase the susceptibility of C3 for inhibition by complement regulators, thus decreasing complement activation. In the same way, a GoF mutation in FH may enhance its ability to regulate C3, thus also decreasing complement activation. For disease, either LoF or GoF mutations may increase either susceptibility or protection.

Following a germline mutation, over generations, the evolutionary mechanisms of both genetic drift and natural selection push the mutated allele to be either lost from the population or increase in frequency and possibly become fixed. The likelihood of evolutionarily deleteriousness is more for LoF mutations than for mutations that either affect non-essential genes or only slightly alter protein function or expression ([Henn et al., 2015](#)). Therefore theoretically deleterious variants will be rare, which has been supported by observations of large proportions of deleterious variants that are indeed rare ([Kryukov et al., 2007](#); [Zhu et al., 2011](#)). However, not all RVs will be deleterious. The average exome contains 7.6 rare non-synonymous variants with $MAF < 0.1\%$ in well-characterised dominant disease genes, without disease ([Lek et al., 2016](#)). For deleterious RVs, their persistence at low frequencies in a population appear to be enabled by either the heterozygous advantage ([Section 2.3.2](#)), population bottlenecks (genetic drift) or rapid growth ([Maher et al., 2012](#)).

2.5 *In silico* predictive tools

In order to assess the functional impact of variants on proteins, a number of *in silico* predictive tools have been developed. Although experimental studies provide the most reliable variant functional analyses, *in silico* computational tools provide a quick and inexpensive theoretically-driven method of variant prediction. Thus, *in silico* tools are particularly useful for analysing large numbers of variants rapidly generated from new next generation sequencing methods. However, they do not always provide reliable and congruent evidence for variants of unknown significance ([Ernst et al., 2018](#); [Kerr et al., 2017](#)). PolyPhen-2 is a program which predicts the possible impact of an amino acid substitution on the structure and function of a human protein by using straightforward physical and comparative considerations. By this, PolyPhen-2 uses both sequence and structure-based predictions, including multiple sequence alignment, accessible surface area and hydrophobic propensity analyses, which feed into a machine-learning program known as a naïve Bayes classifier and result in a functional prediction. For PolyPhen-2

analyses, the user chooses either a HumVar or HumDiv trained dataset, which distinguishes either mutations with drastic effects (RVs in Mendelian diseases) from other human variation, or only mildly deleterious alleles (RVs in complex phenotypes and other uses) from other human variation, respectively ([Adzhubei et al., 2010](#)). 'Sorting Intolerant From Tolerant' (SIFT) is another algorithm which predicts whether an amino acid substitution affects protein function. For an amino acid substitution, SIFT uses sequence homology via protein sequence database searches and multiple sequence alignment to assess its conservation. The output is either 'tolerated' or 'deleterious' ([Kumar et al., 2009](#)). One limitation of SIFT is that it does not analyse protein structure. For variants, both Polyphen-2 and SIFT were better at predicting those that were LoF than those that did not affect function or were GoF ([Min et al., 2016](#)). This difference in sensitivity may be due to LoF variants having more severe consequences than both GoF and benign variants, which likely increases the confidence with which the programs base their predictions for LoF variants. For example, a LoF variant that causes a large change in the physicochemical properties typically leads to protein misfolding. In contrast, GoF (and benign) variants may have a more subtle effect on protein structure thereby resulting in a prediction of lower confidence by the program. In addition, GoF variants are expected to be less common than LoF variants and this is reflected in the variant training datasets for the programs ([Flanagan et al., 2010](#)). Other similar *in silico* tools include PROVEAN (Protein Variation Effect Analyzer), Mutation Assessor, Panther, Combined Annotation Dependent Depletion (CADD), and Condel.

2.6 Terminology of mutations, polymorphisms and variants

For sequence variants in human DNA and protein sequences, the terms 'mutation' and 'polymorphism' were initiated by two papers in 1993 ([Beaudet & Tsui, 1993](#); [Beutler, 1993](#)). Mutation refers to any rare change in the nucleotide sequence which is usually but not always linked to a disease attribute. Mutation may or may not cause phenotypic changes. Polymorphism refers to a variation in DNA sequence that occurs in the general population with a frequency of 1% or higher. Thus, the threshold of 1% was established to distinguish common (polymorphism) from rare (mutation) variants ([Brookes, 1999](#)). For the general population, the more common the polymorphism, the more likely it has a neutral or beneficial effect. However, mutations that were thought to be rare have been found to exceed the 1% frequency threshold ([Auer et al., 2012](#)). In another study, alleles common in one population were frequently not common in another population.

Furthermore, a disease-associated mutation in one population was found to be harmless in another, and vice versa (Myles et al., 2008). Thus, the terms mutation and polymorphism can be independently used to describe the same event with different reference genetic datasets. The distinction between mutation and polymorphism on the basis of their disease-causing capacity is further complicated, e.g. for disease, common variants can be associated with either increased risk or increased protection. Overall, the terms mutation and polymorphism can lead to confusion. In order to resolve this issue, the American College of Medical Genetics and Genomics recommended the term ‘variant’ be used instead with the modifiers: pathogenic, likely pathogenic, uncertain significance, likely benign, or benign (Richards et al., 2015; Karki et al., 2015).

2.7 Sequencing methods

For the investigation of genotype-phenotype relationships, a reference human genome sequence is required. The first human reference genome was assembled by The Human Genome Project in 2001 (Lander et al., 2001). The assembly was based on multiple genomes of anonymous donors in the US. Despite this human reference genome taking 10 years to sequence at first, the sequencing of entire genomes has now become routine in research and medicine and can take 1-2 days or even less. A combination of long-range sequencing and assembly technologies has made highly contiguous whole genome de novo (reference) assemblies possible. The current human reference genome assembly, GRCh38, has a total of 3,088,269,832 nucleotides and was constructed by using Sanger sequencing. Sanger sequencing (first generation) is based on the chain-termination method of sequencing (Sanger et al., 1977) and can produce long, 1000 nucleotide reads. This can make it 10 times more accurate than high throughput short read sequencing, such as next-generation sequencing (second generation) (Guo et al., 2017). Next generation sequencing encompasses a group of highly parallel DNA sequencing technologies that can produce hundreds of thousands or millions of short reads for a low cost and in a short time. When compared to Sanger sequencing, next generation sequencing is more sensitive to low frequency mutations (Arsenic et al., 2015). For *CFH*, due to its sequence homology with the five *CFHRs*, the longer reads obtained from Sanger sequencing are much more accurate than next generation sequencing. By using either Sanger or next generation sequencing methods, certain parts of the genome can be sequenced. Firstly, whole genome sequencing produces the most complete dataset for an individual’s genome, however both the costs and the time needed for computationally

analysing the massive amount of data are great. Despite being made quicker and less expensive by next generation sequencing technologies such as Illumina dye sequencing in 2005, whole genome sequencing still remained expensive. This prompted the development of targeted gene sequencing and whole exome sequencing. In targeted gene sequencing, only the genes of interest are sequenced. By this, the costs are much lower and the coverage is higher, but the detection of disease-associated variants will only succeed if the disease-causing gene is included in the panel ([Sun et al., 2015](#)). In contrast, whole exome sequencing captures only the protein-coding DNA. Whole exome sequencing is cheaper and provides an efficient strategy for the detection of disease-causing variants in proteins, but can miss both major types and regions of disease-causing genomic variation such as structural and intronic variants. At present, the cost of whole genome sequencing and whole exome sequencing have each approached US\$1000 and a few hundred US\$, respectively. This has greatly accelerated the pace of individual human sequencing ([Shendure et al., 2017](#)).

After the genomic DNA fragments are sequenced, they are aligned and merged based on overlapping nucleotides. This leads to contiguous segments or contigs of DNA for which the sequence of bases can be accurately calculated. Multiple contigs are then assembled to form a scaffold with gaps. Multiple scaffolds are then combined to form a chromosome. After this, genomic annotation involves predicting the features of the DNA, such as coding genes, pseudogenes, promoter and regulatory regions, untranslated regions and repeat regions ([Reeves et al., 2009](#)). Advanced computing methods and algorithms are required for these processes. For the detection of copy number variants, either multiplex ligation-dependent probe amplification or array based technologies are used to supplement whole genome sequence, whole exome sequencing and targeted gene sequencing ([Dillon et al., 2018](#)).

Complete genome sequences for humans has led to the development of millions of polymorphic markers. For an individual, a whole genome-based SNP profile can be obtained by a SNP chip. Over the entire human genome, the International HapMap Project has mapped blocks of SNPs based on haplotype analyses ([International HapMap et al., 2007](#)). Haplotypes occur due to a lack of genetic recombination between sites, which strengthens the linkage disequilibrium.

2.7.1 Sequence databases and tools

For nucleotide sequences, three publicly accessible databases hold both eukaryotic and prokaryotic data. These are the nucleotide sequence database of the European Molecular Biology Laboratory (EMBL) at the European Bioinformatics Institute (EBI) ([Kanz et al., 2005](#)), Genbank at The National Center for Biotechnology Information (NCBI) of the National Institutes of Health (NIH) ([Benson et al., 2008](#)) (US) and the DNA Database of Japan (DDJB) ([Kodama et al., 2018](#)) at the National Institute of Genetics in Mishima. For example, a total of 78 eukaryote and prokaryote reference genome sequences are provided by the EMBL Ensembl web-database. For conserved residue analyses via multiple sequence alignment, Ensembl and its tool Compara can be used to retrieve the DNA sequences of protein homologs from numerous different species. Ensembl also includes the prediction of genes, transcripts and peptides ([Flicek et al., 2008](#)). The EBI also provides tools such as Clustal Omega for the multiple sequence alignment of either DNA or protein sequences, and resources such as UniProt ([The UniProt, 2017](#)) for protein sequence and functional annotation. The Reference Sequence (RefSeq) collection by NCBI provides annotated, non-redundant sequences for genomic DNA, transcripts and proteins ([Pruitt et al., 2007](#)). RefSeq genomes are copies of selected assembled genomes from GenBank. As of Release 88, RefSeq contained over 132 million sequences, including 110 million protein sequences, which represented 79,448 organisms ([O'Leary et al., 2016](#)). The RefSeq transcript and protein records are generated by a combination of annotation pipelines, manual curation and other annotated genome resources. The NCBI also provides dbVar and the online version of the Online Mendelian Inheritance in Man database of human-inherited diseases and their associated genes. Online Mendelian Inheritance in Man contains about 18,000 entries and includes data on over 12,000 established gene loci and phenotypes. In addition, the NCBI provides a repository of tools which allows analyses to be performed for several types of gene, protein, and genomic data, such as the Basic Local Alignment Search Tool (BLAST) programs ([Wheeler et al., 2008](#)). The University of California Santa Cruz (USCS) genome browser displays genomic data for human, mouse and other organisms. USCS also provides genomic annotation which was either computed by USCS or contributed to by their collaborators. The genomic annotation is visualised as horizontal tracks ([Kent et al., 2002](#)). ClinVar is a public database of reports of the relationships between human variation and phenotype with supporting evidence. For variation calls in ClinVar, the

level of confidence in their accuracy depends on the supporting evidence, which is variable and provided by the resource ([Landrum et al., 2018](#)).

2.8 Genetic variants and disease

Single-gene or monogenic diseases run in families and can be either dominant or recessive, and either autosomal or sex-linked. For this reason they are known as Mendelian diseases and are usually rare. For monogenic diseases, pedigree analyses of large families with many affected individuals are used to determine the pattern of inheritance. However, sporadic cases of monogenic diseases are also seen. For monogenic disease, the affected gene will likely harbour a deleterious functional variant which increases disease susceptibility and is nearly fully penetrant. For the deleterious variant, the severity of the associated phenotypes and large reduction in fitness prevents it from rising to higher frequencies in populations. It is therefore kept at a very low AF in the population. However, in individual carriers, such a variant could be required but not sufficient to cause the Mendelian disease. By this, modifiers such as additional variants in the genetic background ([Mitchell, 2012](#)), as seen in Huntingdon's disease ([Kearney, 2011](#)), or environmental triggers are required. The likelihood of the variant causing the disease by itself is known as penetrance. For a disease, the penetrance of the variant is calculated by dividing the number of patients that have the variant by the total number of patients. Variants that have lower penetrance might require additional 'hits' to result in the disease phenotype. However, highly penetrant monogenic diseases can still be modified by other genetic variants ([Mitchell, 2012](#)). Dominant variants occur largely as a result of sporadic mutation, whereas recessive variants are maintained at low frequencies by the process of natural selection against homozygotes. In total, approximately 7,000 rare, monogenic diseases are known. Collectively, monogenic diseases are common ([Boycott et al., 2013](#)) and may affect as many as 30 million Europeans ([Dodge et al., 2011](#)). The genetic basis is only known for around half of rare monogenic diseases ([Boycott et al., 2013](#)).

A complex or multifactorial disease is thought to develop due to the combined effects of a large number of genetic variants in an affected individual. For complex disease, no single locus contains alleles that are necessary or sufficient for disease manifestation. This is distinct from a model of genetic heterogeneity in which many different genetic variants are involved in a disease across the whole population, but each

case is caused by a single or a few variants ([Mitchell, 2012](#)). Alongside genetics, non-genetic environmental factors also influence the manifestation of complex disease. Thus, the multifactorial nature of complex diseases means that they do not usually follow Mendelian inheritance in familial segregation analyses. A small fraction (<1% - 7%) of complex diseases are associated with single mutant genes that were transmitted by Mendelian inheritance ([Scheuner et al., 2004](#)), but these cases often have an earlier age of onset with more severe clinical symptoms ([Motulsky, 2006](#)). For environmental factors, their importance in the emergence of disease phenotypes has been shown by the incomplete concordance of the phenotypes of monozygotic twins for many diseases. Many common diseases are complex, such as hypertension, diabetes, Alzheimer's disease and AMD. Complex diseases often have late onset. For each of the multiple variants that contribute to the traits of a complex disease, the penetrance will be lower than for variants that contribute to Mendelian disease traits ([Figure 2.4](#)). For carriers of complex-trait variants, there is often little or no effect on their reproductive fitness. By this, in the population, the complex-trait variants may become more common due to random genetic drift ([Wagner, 2013](#)).

For common complex diseases, the genetic architecture has been modelled by two hypotheses. The “common disease, common variant” hypothesis states that risk alleles for common complex diseases should be common (>5%) ([Figure 2.4](#)) and therefore old and found in multiple human populations rather than being population specific ([Myles et al., 2008](#)). This was coupled with the discovery of several susceptibility variants for common disease that had high MAF ([Bush & Moore, 2012](#)). In keeping with the “common disease, common variant” hypothesis, the “allelic spectrum of disease” is a theoretical perspective that includes low and high penetrance and common and RVs. On the other hand, the “common disease, RV” hypothesis focuses on RVs because, despite the total frequency of variant alleles that contribute to a trait at one locus being moderately high, the frequency of each of the individual variants can be much lower (i.e. rare). This theory was initially based on theoretical simulations of complex disease traits. Further evidence for the “common disease, RV” model was provided by a literature review of genetic association studies ([Wagner, 2013](#)).

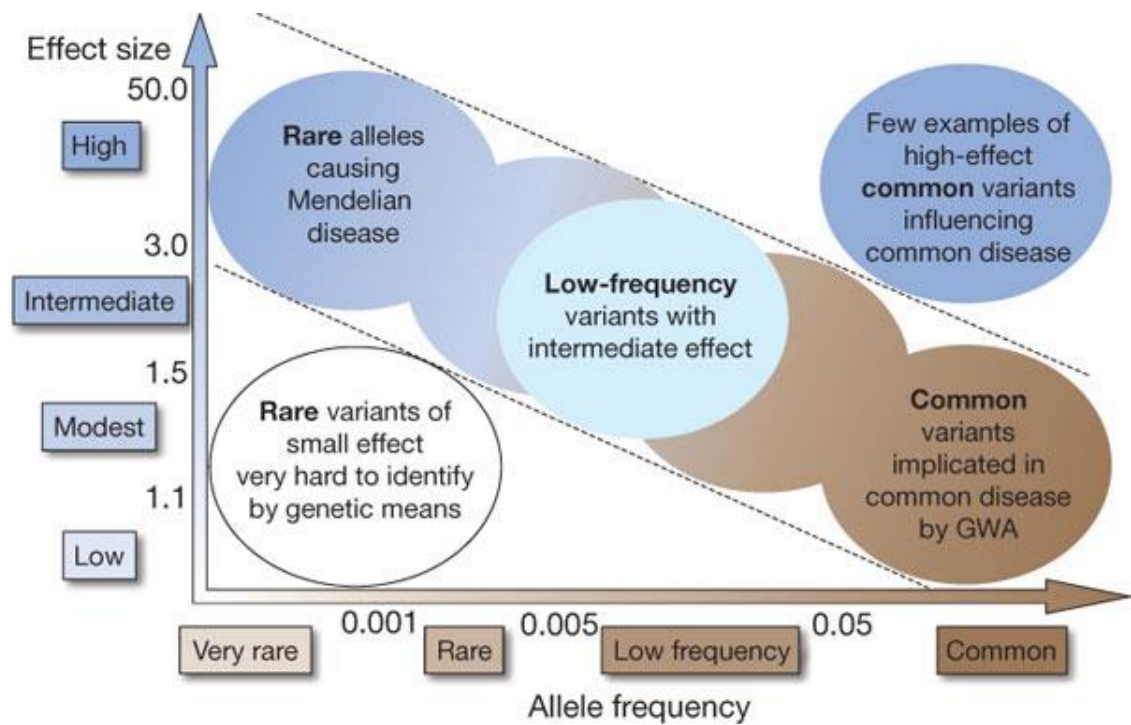


Figure 2.4 Spectrum of disease allele effects. Allele frequency (AF) (x axis) and effect size (y axis) are often used to conceptualise disease associations. Alleles for Mendelian diseases are very rare and highly penetrant with high effect sizes (upper left). Most genome-wide association study (GWAS) findings are of common single nucleotide polymorphisms with low effect sizes (lower right). The bulk of discovered genetic associations lie within the diagonal dotted lines. Sourced from ([Manolio et al., 2009](#)).

2.8.1 Family-based studies

In order to associate a genotype with a disease-related phenotype, either family-based or population-based case-control studies are employed. Family-based studies are most efficient for analysing either rare monogenic diseases or rare sub-phenotypes of common conditions ([Zondervan & Cardon, 2007](#)). For family-based studies, the number of relatives studied can range from two family members to enormous pedigrees. The affected individual is known as the proband. In these families, both linkage and association studies directly evaluate genetic markers. In contrast, other family-based studies that use segregation, aggregation or twins evaluate the potential genetic basis of disease using patterns. For example, segregation analyses allow the mode of inheritance of the disease to be established.

For rare monogenic diseases that have known inheritance patterns in families, linkage analyses are arguably the most powerful method. For linkage analyses, the inheritance of genetic markers spaced evenly throughout the entire genome, such as SNPs, with the disease is evaluated. If the markers and the disease trait co-segregate in families, the disease-causing variants are inferred to be near the markers. Model-based (parametric) linkage analyses are used for large pedigrees, whereas a model-free approach (non-parametric), which is less powerful and based on identity-by-descent estimates, is used for a pair of affected siblings. By these linkage methods, *a priori* pathophysiological hypotheses are not needed. Overall, linkage analyses do not aim to identify the causal gene but facilitate the process of directly identifying disease variants. An advantage of family-based studies is that population stratification issues are bypassed because of the common genetic background between individuals ([Schnell & Witte, 2008](#)). Furthermore, families with multiple disease cases are enriched for causal variants which leads to increased statistical power. A disadvantage is that it can be difficult to accumulate large enough samples of well-characterised families for analyses ([Evangelou et al., 2006](#)), and linkage analyses are not always applicable to small family sizes ([Barc & Koopmann, 2011](#)). In addition, for sporadic cases of rare diseases, for which there is no previous family history of the disorder, family-based studies are limited. The power of linkage analyses is greatly reduced by incomplete penetrance. In order to overcome this loss of power, multiple families consisting of small numbers of affected individuals may be analysed via aggregation. However, in the presence of locus heterogeneity, the

aggregation may result in a further loss of power, because different families will have different genes associated with the disease (Guo et al., 2016).

For complex and common diseases, family-based studies are not well suited unless the correlation between genotype and phenotype is very robust. Instead, population-based case-control studies are utilised.

2.8.2 Population-based case-control studies

For population-based case-control studies, individuals with the disease-associated trait (cases) and individuals without (controls) are selected from the population. In order to test whether a genetic variant of interest is associated with the trait, the variant AF is calculated for each group and subsequently compared. If the variant is significantly more frequent in cases than controls, the variant is statistically associated with the disease. This method can be applied to either rare or common variants. However, the sample sizes needed to detect RVs ($AF < 1\%$) can be unfeasibly large (Zondervan & Cardon, 2004). In order to generate the genetic data for population-based case-control studies, all individuals in both populations should be sequenced by using the same method. This can be either targeted to specific genes, from *a priori* knowledge of the genes implicated in the disease (candidate gene method), or non-hypothesis based such as for genome-wide association studies (GWAS). Relative risk and odds ratio statistics can also be calculated.

For these population-based case-control studies, despite the high statistical power from variant enrichment in cases, there is often an increased susceptibility to false positives due to the absence of data for the genetic backgrounds of cases and controls (Altshuler et al., 2008). Thus, spurious associations can be detected if cases and controls are sourced from different populations for which the AFs vary (Schnell & Witte, 2008). Such unaccounted population admixture can lead to confounding results. Ethnicity matching is thus very important in order to limit potential confounding by genetic ancestry.

For GWAS, the aim is to correlate SNPs with a disease status or trait variation by comparing the AF of the variants between cases and controls (van der Sijde et al., 2014). Genotyping panels are typically used and can provide 300,000 – 1 million SNPs. The decreasing costs and increasing SNP density of standard genotyping panels has shifted

the focus of attention from candidate gene approaches towards whole genome based GWAS ([International HapMap et al., 2007](#)). In general, SNPs identified in GWAS that associate with a phenotype are unlikely to be causal variants but proxies. By this, they are likely to be associated via linkage disequilibrium with the true but currently unknown causal variants. GWAS can identify genes which would not have been considered *a priori* to be good candidates for the disease. For example, GWAS in humans have been successful for identifying genetic regions that are important in the development of complex and common diseases such as diabetes ([Sladek et al., 2007](#)), Crohn's disease ([Barrett et al., 2008](#)) and other autoimmune and genetic diseases. One of the most successful applications of GWAS was for common variants in AMD in 2005 ([Klein et al., 2005](#)), which triggered numerous more detailed studies on AMD ([Black & Clark, 2016](#)). One disadvantage of GWAS is that due to the number of independent tests performed, the statistical power is often low.

However, for many complex common diseases, common variants identified by GWAS only explain a small proportion of the total genetic variance to the trait variance. For common diseases, this 'missing heritability' is thought to comprise either many common variants of small effect sizes ("common disease, common variant") which are not fully identified by underpowered GWAS, RVs ("common disease, RV") ([Wagner, 2013](#)) or environmental factors. For AMD in 2011, a high-throughput GWAS using *CFH* genotyping identified a haplotype of high penetrance which lead to the association of a RV (c.3628C>T; p.Arg1210Cys) after statistical analyses. Following this, and with the use of functional data, it was concluded that AMD-risk is likely driven by rare *CFH* LoF alleles ([Raychaudhuri et al., 2011](#)), in addition to common variants.

2.8.3 Allele frequency data for rare variants

For rare Mendelian diseases, such as aHUS and C3G, candidate RVs are typically identified by family-based or small case-control studies. For the clinical interpretation of candidate variants, the rarity of a variant is a prerequisite for pathogenicity. For a variant, the pathogenicity generally decreases as the AF increases ([Kobayashi et al., 2017](#)) ([Section 2.4.3](#)). Thus, a variant expected to be causative for a rare Mendelian disease will not be frequent in unselected individuals (without the disease). The filtering of candidate variants by their AFs in unselected individuals is a key step in any pipeline for the discovery of causal variants ([Lek et al., 2016](#)).

Next generation sequencing methods combined with the development of computer software have allowed a vast amount of sequence reads to be generated, processed, quality controlled and called. By this, RVs can be detected from datasets of sequences from thousands of individuals or more. An example of this is The 1000 Genomes Project (1000GP) ([Genomes Project et al., 2010](#)), which sequenced more than 2,500 healthy individuals from 26 populations by using low-depth whole genome sequencing combined with imputation methods for high-quality variant calls. Due to the difficulty in interpreting non-coding genomic DNA (98%), whole exome sequencing was developed in order to capture the exons only ([Section 2.7](#)). An example of whole exome sequencing for the identification of RVs is the National Heart, Lung and Blood Institute (NHBLI) Exome Sequence Project (ESP) Exome Variant Server (EVS) which sequenced the exomes of 6,515 individuals ([Fu et al., 2013](#); [Tennessen et al., 2012](#)). Such whole exome sequencing has identified the genetic basis of many rare monogenic diseases ([Gilissen et al., 2011](#)). The ExAC is a large protein-coding human variation catalogue of aggregated sequence data from 60,706 individuals of different ethnicities ([Table 2.5](#)) and cohorts ([Table 2.6](#)). The ExAC study identified more than 7.4 million new genetic variants of high confidence. The ExAC is the first dataset with high enough power to identify very RV AFs in the range of 0.01% – 0.0001% ([Walsh et al., 2016](#)).

The ExAC, 1000GP and EVS are large genetic variant datasets of unselected individuals without rare Mendelian diseases. By this, they can be used as reference datasets for the AF-based interpretation of candidate variants, with caveats such as differences between both the sequencing methods and ethnicity of participants. Population stratification can be used to overcome the latter. For a variant to be classified as “benign”, an “allele frequency greater than expected for disorder” in the reference dataset has been recommended as strong evidence by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology in their joint consensus on the interpretation of sequence variants ([Richards et al., 2015](#)). The reference dataset AF cut-off for RVs is typically below 1% and differs depending on the disease of interest and the variant penetrance. Based on analyses of the ExAC data, it was concluded that a variant identified in ExAC with a global AF greater than 0.01% is too common for a disease of Mendelian inheritance. As an alternative to using theoretical incidence and penetrance methods, the AF cut-off for a disease can be determined by analysis of the pathogenic variant burden ([Kobayashi et al., 2017](#)).

Table 2.5 Cohorts represented in the ExAC data

Consortium/Cohort	Number of Samples
1000 Genomes	1,851
Bulgarian Trios	461
GoT2D (Type 2 diabetes)	2,502
Inflammatory Bowel Disease	1,675
Myocardial Infarction Genetics Consortium	14,622
NHLBI-GO Exome Sequencing Project (ESP)	3,936
National Institute of Mental Health (NIMH) Controls	364
SIGMA-T2D (Type 2 diabetes)	3,845
Sequencing in Suomi (SISu)	948
Swedish Schizophrenia & Bipolar Studies	12,119
T2D-GENES (Type 2 diabetes)	8,980
Schizophrenia Trios from Taiwan	1,505
The Cancer Genome Atlas (TCGA)	7,601
Tourette Syndrome Association International Consortium for Genomics (TSAICG)	297
Total	60,706

Table 2.6 Populations represented in the ExAC data

Population	Male Samples	Female Samples	Total
African/African American	1,888	3,315	5,203
Latino	2,254	3,535	5,789
East Asian	2,016	2,311	4,327
Finnish	2,084	1,223	3,307
Non-Finnish European	18,740	14,630	33,370
South Asian	6,387	1,869	8,256
Other	275	179	454
Total	33,644	27,062	60,706

Despite this, some recent studies have shown that the rarity of a variant in the reference dataset is not always a reliable indicator of pathogenicity for a Mendelian disease. In the ExAC dataset, which does not contain rare Mendelian disease cases ([Table 2.6](#)), 192 variants reported to cause rare Mendelian disease were found at high frequencies. The ExAC also showed that the average participant had approximately 54 variants that were previously classified as casual for rare disease ([Lek et al., 2016](#)). It has also been estimated that each healthy genome is heterozygous for 50-100 variants that have been classified as causing inherited disorders in the Human Gene Mutation Database. For the 1000GP, analyses found that each healthy genome contains approximately 100 high-confidence functional LoF variants, of which 20% were homozygous ([MacArthur et al., 2012](#)). These LoF variants include premature stop codons, splice site disruptions and frameshifts ([Genomes Project et al., 2010](#)). These LoF variants are mostly rare which indicates the act of negative selection ([Wagner, 2013](#)). These occurrences of deleterious LoF variants in healthy people implies that some functional variation may be tolerated by genes. By this, the genes most likely to influence disease are those that are the most intolerant of functional variation. In keeping with this, genes responsible for Mendelian diseases show significant intolerance ([Petrovski et al., 2013](#)). For a gene, functional tolerance can be gained via a number of mechanisms. In summary, for analysing RV pathogenicity in disease, AF analyses are essential for support but cannot be used as the sole evidence.

In GWAS, low-frequency (rare) variants are not tagged by conventional genome-wide genotyping arrays (or chips). However, some specialised genotyping chips for RVs have recently been developed based on genomic co-ordinates that are identified by sequencing projects such as 1000GP and EVS. The chips can test many exonic variants at modest costs. Compared to sequencing, RV chips are computationally simpler to analyse but will miss a large amount of very rare genetic variation. Another drawback of RV chips is that they are based on European datasets and so may not reflect genetic variation in other populations ([Auer & Lettre, 2015](#)).

2.9 Methods: rare variant assessment guidelines for rare genetic diseases

For variant pathogenicity classification, the American College of Medical Genetics and Genomics recommends using criteria based on a combination of population, computational, functional and segregation data ([Richards et al., 2015](#)). Based on these

criteria, for aHUS and C3G, the categorisation of genetic variants has recently been standardised ([Goodship et al., 2017](#)) ([Table 2.7](#)). For variant pathogenicity for aHUS and C3G, a reference AF cut-off of <0.1% is recommended. Reference databases currently available include 1000GP, EVS and the ExAC. Overall, the variant is assessed by using these AF analyses in combination with the literature, supporting functional data, the zygosity, *in silico* predictions and the location of the variant within the protein. The opinion of an expert laboratory in aHUS or C3G is also required to determine the disease impact of particular genetic variants. A more stringent RV AF cut-off of <0.01% is also applicable to rare diseases of Mendelian inheritance ([Kobayashi et al., 2017](#)). Compared to <0.1%, the <0.01% AF cut-off restricts the pathogenicity inclusion criteria to rarer variants. This decreases the likelihood of incorrectly including a non-pathogenic variant (background variation), but may also exclude potentially pathogenic variants of less penetrance (not as rare). For describing genetic variants, a standard gene variant nomenclature is maintained by the Human Genome Variation Society.

For clinical genetic testing of aHUS and C3G patients, the minimum set of genes that should be screened includes *CFH*, *CD46*, *CFI*, *C3*, *CFB*, *THBD*, *CFHR1*, *CFHR5* and *DGKE*. For aHUS only, genotyping for the risk haplotypes *CFH*-H3 and *CD46*_{ggaac} is also carried out. In order to detect copy number variants, hybrid genes and other complex rearrangements in the *CFH/CFHR* genomic region, suitable technologies such as multiplex ligation-dependent probe amplification are used ([Goodship et al., 2017](#)). In addition to genetic analyses, plasma complement components, expression of CD46 on peripheral blood mononuclear cells and anti-FH autoantibodies are also measured. However, plasma FH level measurements may not allow detection of heterozygous FH deficiency states.

2.10 Methods: rare variant burden tests

For assessing the association of genes with a disease, RV burden tests can be used. RV burden tests allow the rare variation of a gene for a disease cohort to be compared with large reference datasets of many unselected individuals which represent background rare variation. By using large reference datasets, the power of the study is increased. However, this also depends on the size of the disease dataset. RV burden tests are useful for verifying the results of family-based and small-scale case-control studies for rare diseases, such as for aHUS and C3G. This is because these smaller-scale studies are

Table 2.7 Categorisation of the genetic variants for aHUS and C3G

Category	Criteria
Pathogenic	Novel or rare (MAF <0.1%) variant reported to cause disease in literature; supporting functional data indicating the variant affects protein function or expression; zygosity fits inheritance pattern.
Likely pathogenic	Novel or rare (MAF <0.1%) variants that change protein sequence or affect splicing and with highly deleterious effects by in silico predictions but without functional data; found in disease-related functional domains; zygosity fits inheritance pattern.
Uncertain significance	Novel or rare (MAF <0.1%) variants that change protein sequence or affect splicing with no known functional data; uncertain deleterious effects by in silico prediction.
Likely benign	0.1% <MAF <1%; not in a mutational hotspot; no functional data; in silico predicted benign.
Benign	MAF >1%; variant reported as not-disease associated; functional data suggest benign; benign by in silico prediction.

Sourced from ([Goodship et al., 2017](#))

limited by both genetic factors which are heterogeneous and incompletely penetrant, and diseases for which sporadic cases are often observed ([Section 2.8.1](#)). All of these limiting factors decrease the power of familial studies for detecting gene associations.

For RV burden tests, RVs are collapsed into genetic scores and tested for association with the disease trait. For a gene, this determines whether the aggregated frequency of rare variation identified for the disease cohort is significantly different from that expected for individuals without disease. For the case individuals, RV burden tests are typically applied to unrelated probands. For the reference dataset, AF data from large population-based datasets such as the ExAC can be extracted. In a recent study, the use of population-based control subjects rather than disease-free control subjects showed negligible effects on power ([Guo et al., 2016](#)).

Burden tests are powerful when a large proportion of the variants are causal and the effects are in the same direction. However, burden tests lose power when either a mixture of both trait-increasing and trait-decreasing variants are present, or if only a small fraction of the variants are casual ([Lee et al., 2014](#)). In the RV burden test, the qualifying variants in control subjects represent a background rate of variation for each gene. However, if there is incomplete penetrance, unaffected control individuals may also harbour pathogenic variants. By this, some of this background variation may correspond to incompletely penetrant pathogenic variants that have not yet manifested as disease, as well as potentially misclassified benign variants. Thus, the RV burden test is based on a simplifying assumption that the frequencies of rare benign variants in both cases and the reference dataset are equivalent, and that the frequency of pathogenic variants in the reference dataset is sufficiently low ([Walsh et al., 2016](#)). This is facilitated by including only protein-altering variants and applying the same AF filter to both cases and controls. In addition, in order to reduce false positives, the control cohort can be matched to the case samples for both ancestry and technical factors such as the depth of sequencing coverage in each gene ([Guo et al., 2016](#)). For a burden test, the data is presented in a contingency table and either the χ^2 or the Fisher's exact test of independence can be used for significance testing. Both of these tests assess independence between two variables when the comparing groups are independent and not correlated. The χ^2 test applies an approximation which assumes the sample is large, whilst the Fisher's exact test runs an exact procedure which is more accurate for small-sized samples ([Kim, 2017](#)).

For statistically summarizing the evidence for association between a disease and a gene (or genetic variant), the most convenient method is significance testing with appropriate correction for multiple testing. For significance testing, the P value is defined as the probability of obtaining a value (T), which follows a known probability distribution, that is at least as extreme as that of the actual sample (t), when the null hypothesis (H_0) is true (Fisher, 1925; Neyman & Pearson, 1933). If P is smaller than a pre-set “significance” threshold α then H_0 is rejected and the result is considered to be significant. α can be corrected for multiple testing by using the Bonferroni correction. This divides α by the number of independent tests carried out (Armstrong, 2014). The range of values of the test statistic T that would lead to the rejection of H_0 is known as the critical region of the test. For the test, the probability of rejecting H_0 when it is actually true (a false negative) is known as a Type 1 error and is equal to α . Thus, the more relaxed (greater) the significance level, the more likely there will be a Type 1 error. The probability of failing to reject H_0 when it is actually false (a false positive) is known as a Type 2 error and denoted as β . The statistical power of the test, $(1 - \beta)$, is defined as the probability of correctly rejecting H_0 when it is actually false (true negative) and a true association is present. For the study, the power calculation is used to ensure realistic and meaningful results are possible (Sham & Purcell, 2014) and is often set to 80%. This means that if a difference exists, there will be a 20% chance of a Type 2 error. The power of a study is increased by sample size. The *Altman nomogram* can be used to calculate the sample size needed for 80% power by inputting the standardized difference and α level (Columb & Atkinson, 2016). For categorical data expressed as proportions p_1 and p_2 for two groups, the standardised difference is given by:

$$\text{Standardised difference} = \frac{p_1 - p_2}{\sqrt{p(1 - p)}} \quad (2.6)$$

2.11 Overview of aHUS, C3G and AMD genetics

aHUS and C3G are two ultra-rare diseases that both involve genetic predisposition for complement AP dysregulation and a triggering event. However, between aHUS and C3G, both the molecular mechanisms and clinical presentations differ (Chapter 1, Sections 1.7.1 and 1.7.2). For both aHUS and C3G, a number of predisposing genetic variants have been identified by familial and small-scale case-control cohort studies, with

subsequent functional analyses. For C3G, the genetic understanding is not yet comparable to that of aHUS ([Goodship et al., 2017](#)).

For aHUS, some variants have strong links based on familial studies. For example, the screening of 25 individuals with aHUS within three families showed that all carry a pathogenic *CFH* variant (c.3643C>G; p. Arg1215Gly). However, the penetrance of this strongly linked variant in aHUS is ~50% (incomplete) and is determined by the *CFH* and *CD46* haplotype, additional RVs, age and a trigger ([Sansbury et al., 2014](#)). Overall, the investigation of individuals with aHUS for genetic susceptibility factors has become increasingly complex because both the number and nature of reported genetic defects has expanded ([Goicoechea de Jorge & Pickering, 2010](#)). Presently, in ~60% of aHUS patients, one or more genetic abnormalities in complement AP or related genes have been detected. These mostly drive AP dysregulation at the endothelial cell surface ([Goodship et al., 2017](#)). For genetic aHUS, most cases are heterozygous ([Nester et al., 2015](#); [Rodriguez et al., 2014](#)), and are attributed to the genes *CFH*, followed by *CD46*, *C3* and *CFI*, and *CFB*. In addition, rare copy number variation in the genomic region which encompasses the *CFHRs* are risk factors for aHUS. For example, the *CFH/CFHR1* and *CFH/CFHR3* hybrid genes. For aHUS, RVs in the thrombosis-related genes thrombomodulin (*THBD*) and plasminogen (*PLG*) have also been identified ([Osborne et al., 2018a](#)). For aHUS, in addition to the initial association of complement genes by using linkage and candidate gene analyses, whole exome sequencing followed by co-segregation studies identified a highly significant association with recessive RVs in the non-complement gene *DGKE* ([Lemaire et al., 2013](#)). For rare diseases, this demonstrates the importance of genome and exome-wide methods in identifying genes for which an association could not have been predicted from previous immunological pathway data.

For C3G, approximately 20% of cases are associated with predisposing RVs in the complement genes. Thus, the vast majority of cases are associated with C3 nephritic factor or autoantibodies. These genetic C3G cases are either sporadic or familial and can be inherited in either autosomal recessive or dominant forms. Familial C3G is most often linked to highly penetrant heterozygous copy number variation in *CFHR1–5* genes, such as *CFHR5* nephropathy ([Gale et al., 2010](#)), as well as homozygous *CFH* deficiency and heterozygous GoF mutation in *C3* ([Osborne et al., 2018a](#)).

AMD is distinct from aHUS and C3G in that it is a common and complex disease, but similar to aHUS and C3G in that it is associated with both complement AP dysregulation and genetic variants. Thus, GWAS have revealed significant statistical associations between AMD and the complement genes *CFH*, *CFHR1*, *CFHR3*, *CFB*, *C2* and *C3* (Anderson et al., 2010). As well as increasing the risk for AMD, common variants have also been associated with reduced activation of the AP thus reduced susceptibility to AMD for both *CFB* (Montes et al., 2009) and *CFH* (Tortajada et al., 2009). For AMD, one of the most significant statistical associations exists for the common *CFH* variant c.1204T>C which encodes the missense change p.Tyr402His. In FH, p.Tyr402His weakened the binding of FH SCR-7 to eye tissue-specific HS only (Clark et al., 2013), which is relevant to AMD. For common variants in complement, each individual has a functional ‘complotype’, which may influence susceptibility to AP-driven disease such as aHUS, C3G and AMD (Heurich et al., 2011).

2.12 Methods: genetic variant web-databases

Currently, the most reliable method of assessing whether a genetic variant is causative or a risk factor for disease consists of a combination of statistical association, by using case-control or family studies, functional assays and *in silico*-based conservation and predictive analyses. By this, large amounts of complex data may exist for each variant. Furthermore, for both rare and common diseases, one case may have multiple identified variants of different frequencies in multiple genes. At the population level, the disease may be associated with a number of different genes. For genetic variants within these contexts, databases allow the storage, organisation and analysis of the data and complex relationships. Databases provide critical resources for the clinical interpretation of variants identified in patients with rare diseases (Lek et al., 2016). For example, the interactive database for human coagulation factor IX assisted in making judgments about hemophilia B patient management, as well as providing insights into the molecular mechanisms of the disease (Rallapalli et al., 2013).

Databases are organised collections of related data tables. A database server stores, retrieves and allows quick, simultaneous access to data from different users and web servers. Databases allow the management of complex relationships between related data tables including one-to-one, one-to-many, and many-to-many. In order to manage databases, a database management system provides a suite of specially designed

computer applications which enables the user, other databases and applications to interact with the data. The relational database management system is an improved version of the hierarchal and network models of database management systems that allows any table to be accessed directly and not via parent objects. For relational database management systems, an object-orientated system is where reusable chunks of applications (objects) are stored in the database, and instructions (commands) enable the object to be used. An object-relational database system thus combines the features of object-orientated storage and relational combinations of the data tables.

Structured Query Language (SQL) was developed in the 1970s for the creation, manipulation and retrieval of data in databases. MySQL is a widely used database server or relational database management system which relies upon SQL. By the use of SQL, data in tables or whole databases can be defined, removed, modified, inserted, deleted, and retrieved, and administrative tasks such as user authorisation and monitoring, security, backup and recovery can be completed. Many application programming interfaces are able to interact with MySQL. SQL is compatible with web development languages such as the PHP: Hypertext Preprocessor (PHP) server-side scripting language thus allowing client/server interactions ([Williams & Lane, 2002](#)). Usually, PHP is embedded or combined with the Hypertext Markup Language (HTML) of a web page. When the page is requested, the web server executes the PHP script and the result is presented via substituting the webpage. By using PHP, dynamic pages can be created with content derived from either user input or a database. For PHP, many libraries exist for fast and customised access to relational database management systems. In order to query a MySQL server and produce the result in a HTML format for display in a web browser, many PHP scripting techniques are available ([Table 2.8](#)).

When made into a web-database application, a database can become a scientific and medical community-based research tool and also securely connect to other web-databases, analytical tools and APIs for in-depth analyses. In order to run dynamic database-driven websites, a stack of free software programs that include an Apache server, MySQL and PHP for either Windows or Linux-based operating systems are commonly used. For a database, a schema provides a complete overview or plan of the data table organisation and relationships. Thus a schema shows the tables with all of their columns, the primary (identifying column) and foreign (identifying column in another table) keys. In order to design a schema, an Enhanced Entity Relationship model is

Table 2.8 PHP and SQL commands for a web-database application

Command	Language	Use example in web-database application	Example command from my Complement database (Chapter 4)
SELECT; FROM; WHERE	SQL	Select columns (* for all) from a table where certain conditions are fulfilled (e.g. column data must be equal to certain mutation ID).	\$q="SELECT * FROM mutations WHERE mut_id=\"\".\$line['mutid'].\"\"";
COUNT; GROUP BY	SQL	Select the number of patients where two conditions are satisfied (e.g. patient ID equal to PHP variable \$string). Count the number of patients for each different combination of the lab ID, zygosity, disease and source (by using GROUP BY).	\$q2="SELECT COUNT(pat_id) as countpatid, zygosity, pat_lab_id, pat_cond_cat, source, ref_id_pat from patient WHERE pat_id IN (\".\$string.\") \$pacond GROUP BY pat_lab_id, zygosity, pat_cond_cat, source";
mysqli_connect	PHP	Connect to database.	\$link = mysqli_connect(\$db_host, \$db_user, \$db_pwd, \$dbname);
mysqli_query	PHP	Query database.	\$resultTotal = mysqli_query(\$link, \$sql);
mysqli_num_rows	PHP	Count the number of rows retrieved.	\$total_records = mysqli_num_rows(\$resultTotal);
mysqli_error	PHP	If query fails, display error message.	mysqli_error("Could not execute query");
mysqli_fetch_array	PHP	Fetch every row as an array, for which each column can then be specified.	while(\$rowPatient=mysqli_fetch_array(\$resultPatient)){...
round	PHP	Round a number up to a specified amount of decimal places.	\$sexacmeanAN = round(\$rowmeanAN['suman']/\$noexacanrows,0);
pow	PHP	Exponential expression.	\$schisq = (pow((ABS(\$obs-\$exp)-0.5),2))/ \$exp;
ABS	PHP	Absolute value.	\$schisq = (pow((ABS(\$obs-\$exp)-0.5),2))/ \$exp;
for	PHP	For each variable (data) from each row of a column in one database table...	for(\$t=0; \$t<count(\$arraytype); \$t++){...
array_search	PHP	Search array for given value.	\$genematch = array_search(\$geneforsearch, \$arrayallgenes);
strlen	PHP	Return length of a string.	if(strlen(\$geneforsearch) == strlen(\$arrayallgenes [\$countaa])){...

commonly used for the design and conceptual presentation. University College London provide secure web and MySQL (database) servers that can be accessed by Linux and the web, respectively.

Database security against intruders must be prioritised, especially for databases that contain sensitive or patient identifiable data and are connected to the internet. Before being used in a web-database, sensitive or patient identifiable data must be anonymised in accordance with ethics procedures in order to reduce the risk of patient identification. In order to certify that the data has been anonymised and is safe for a web-database, an application can be submitted to the Information Technology support team for the School of Life and Medical Science at University College London. One type of malicious attack via the web is code injection, in which the attackers use vulnerabilities in the code to either redirect users to other web locations or inject new code that is harmful to the web-database (Welling & Thomson, 2009). By this, the database may be destroyed, modified or copied. In order to prevent this sort of malicious behaviour, the programming of the web-database application is sanitised to eliminate any vulnerable (accessible) points. This can be checked by using the Python-based open source penetration testing tool sqlmap (<http://sqlmap.org/>), which automates and simulates the process of detecting and exploiting any SQL injection flaws in the database. Thus, using PHP, each query is only transmitted to the database if the variables correspond to pre-set known variables. One example of a vulnerability is the web address line which can be used to transmit variables (from PHP) for querying the MySQL database using SQL.

2.12.1 Genetic variant web-databases for aHUS and C3G

For the analysis of genetic variants in aHUS and related diseases, an interactive FH-HUS web-database was set up at University College London in 2006 for both clinicians and researchers to use (Saunders et al., 2007; Saunders et al., 2006; Saunders & Perkins, 2006)(<http://www.fh-hus.org>). At first, the FH-HUS database contained data on published complement genetic variants in *CFH* only (Warwicker et al., 1998), and was then expanded to include the other complement genes *CFI*, *C3*, *MCP* and *CFB* following literature reports (Rodriguez et al., 2014). The FH-HUS database also includes genetic variant data for other diseases involved in complement AP dysregulation, such as MPGN and AMD. By 2014, there were 193 *CFH*, 130 *CFI*, 86 *CD46* and 64 *C3* variants for aHUS, C3G and other diseases. For each variant, the FH-HUS database features structural

analyses, based on whether it falls within a known ligand-binding region and the residue physicochemical differences. These structural analyses are facilitated by both side chain accessibility calculations and a three-dimensional view of the residue involved in the variant via a Jmol applet. For each of the four proteins, both a statistical analysis of the frequency of variants in each domain and the latest structural models are also provided.

For aHUS, C3G, AMD and other genetic diseases, the current challenge for researchers is to separate disease-associated genetic variants from the broader background of variants present in all human genomes that are rare, potentially functional, but not pathogenic for the disease (Vieira-Martins et al., 2016). This is one of the major research questions addressed in this PhD thesis. This background variation can now be well represented, with caveats, by the ExAC, the 1000GP and EVS datasets, which were generated by the recent onset of sequencing and computing advances. Thus, despite the extensive utility and high popularity of the FH-HUS database, a major omission was the inclusion of variant AF data from both reference and disease (aHUS and C3G) datasets. AF data is essential for assessing variant pathogenicity (Section 2.8.2 and 2.8.3) and can be used to verify disease-gene associations (Section 2.10). In addition, by 2014, the number of genes associated with aHUS and/or C3G increased to 13 (*CFH*, *C3*, *CFI*, *CD46*, *CFB*, *DGKE*, *CFHR5*, *CFHR1*, *CFHR3*, *CFHR4*, *PLG*, *CFP*, *THBD*), and many cases have more than one variant that requires analysis. Despite their AP regulatory activities, neither *CRI* (Schramm et al., 2015) nor *DAF* (Kavanagh et al., 2007) are major susceptibility genes for aHUS or C3G. For *CR1*, this may be because it is not expressed on endothelial cells (Roumenina et al., 2009). Since the previous FH-HUS database, new methods for assessing variant pathogenicity, such as conservation analyses by using multiple sequence alignment and *in silico* tools, have also been developed. Overall, the FH-HUS database required an extensive update to include the latest genetic variant data for both aHUS and C3G and facilitate their pathogenicity analyses. This update included, for each gene, extracting the relevant AF data from large reference dataset web-servers and their subsequent presentation on the web-database for both pathogenicity and RV burden analyses (Chapter 4). It also included a new solution structure for FH (Chapter 5) in order to more accurately predict the functional impact of disease-associated *CFH* variants, such as the common AMD-risk variant p.Tyr402His and rare missense variants associated with aHUS and C3G (Chapter 6).

In order to calculate variant AFs for each of the aHUS and C3G populations, a substantial number of cases are required. For this, data from multiple clinical centres for aHUS and C3G can be aggregated. These data include the number of patients that had the variant, the zygosity and the number of patients screened for the gene in total. In order to provide these data for this PhD thesis, a collaboration with six clinical centres for aHUS and C3G was established. By this, new unpublished variant data was also obtained. Prior to this, the largest registry for aHUS was The Global aHUS Registry which was established in April 2012 and had 826 enrolled patients by June 2015. Thus, in [Chapter 4](#) of this PhD thesis, I created a new Database of Complement Gene Variants and used this to analyse AF data and verify the association of RVs in the complement and related genes with aHUS and C3G. These RV data were then used in [Chapter 6](#) to assess whether the structural location of a rare missense variant in the complement proteins can predict the outcome of aHUS or C3G in patients.

Chapter Three

Protein structure and dynamics

This chapter describes the theory, including the roles of amino acid residues, in protein structure, function and dynamics, and how these may be perturbed in disease. It describes methodologies for predicting the structures of flexible proteins in solution by using a combination of biophysics and bioinformatics, such as: low-resolution small angle X-ray scattering, homology modelling, molecular dynamics and Monte Carlo-based modelling. These methods are either used or referred to in my results [Chapter 5](#) for studying the structure of complement factor H and the effect of the disease-risk p.Tyr402His variant on its overall domain arrangement. The protein structure and function theory described here is also used in my results [Chapter 6](#) to address whether the structural locations of rare missense variants in the complement proteins can predict complement disease phenotypes.

3.1 Protein structure and function

For a coding genetic variant that is statistically associated with a disease-related phenotype, the normal function of the encoded protein is expected to be dysfunctional. By this, the dysfunctional protein leads to an imbalance in the biochemical pathway and system. By causing an imbalance, certain components of the pathway will either increase or decrease, due to the inability of the dysfunctional protein to carry out its normal biochemical function. For example, in the complement AP, FB reacts with C3b to form the AP convertase C3bBb and also provides the C3bBb catalytic site for the proteolytic activation of C3 to C3b. For FB, the GoF variant p.Phe286Leu identified in aHUS was found to either increase the resistance of FB to regulation, or increase the formation of C3bBb. This is thought to be due to a decrease in the restriction of mobility for a residue at the C3b–Bb interface ([Goicoechea de Jorge et al., 2007](#)). By this, the levels of C3b increase, thereby predisposing for complement dysregulation and/or the abnormal deposition of complement activator components ([Chapter 1, Sections 1.7.1 – 1.7.3](#)). Predisposition is also driven by LoF variants in the AP regulators such as FH. For LoF variants in FH, the affected residues are often involved in binding to either host cell surfaces, C3b or FI, or involved in stabilising the FH SCR domains. For example, for FH, aHUS-associated missense variants often affect Cys residues that stabilise the SCR domain via disulphide bonds ([Rodriguez et al., 2014](#)) ([Section 3.3.8](#)). For aHUS and C3G, severe symptoms rapidly develop, and this is thought to be due to both strong predisposing variants that are rare in the general population and a triggering event. On the other hand, for the development of AMD, which is most common in the elderly

population, subtler age-related changes together with the presence of risk variants have been proposed to cause subtle symptoms that gradually worsen over a long period of time. For example, for AMD, the age-related decrease in the sulfation of the glycosaminoglycan HS structures combined with the FH p.Tyr402His risk variant have been proposed to reduce FH binding over time. Thus, p.Tyr402His was found to weaken the binding of FH SCR-7 to eye-specific HS at Bruch's membrane, which leads to increased chronic local inflammation at the RPE/choroid interface (Clark et al., 2013) (Chapter 1, Section 1.7.3). This was concluded experimentally by comparing the distribution of fluorescently labelled FH Tyr402 and His402 along Bruch's membrane (Clark et al., 2010). However, the full length structure of FH and how it is affected by p.Tyr402His is currently unknown. In order to study the molecular mechanisms of such disease susceptibility variants at the protein level, the physical properties of the protein, such as structure, stability, and dynamics, are typically characterised.

3.1.1 Protein function

Proteins are complex macromolecules that each execute a specific function at either the biochemical, cellular or phenotypic level. For many proteins, biochemical functions involve molecular interactions which can drive activities such as enzyme mechanisms, ligand binding, cell surface adhesion, membrane transport, the building of large structures (e.g. ribosomes), DNA reading and transcription control. Such biochemical functions are essential for life. The spatial arrangements of the residues in the protein structure provide optimal chemistry and configurational space for molecular interactions. For a protein, the relationship between structure and function is complex in that one protein-folding topology can support more than one function, and one function can be associated with more than one fold (Orengo et al., 2003). In addition, each domain type usually has one or more distinct biochemical function, and the majority of proteins are multi-domain, e.g. up to 80% in eukaryotes (Apic et al., 2001). Within genomes, protein domains have frequently been shuffled and recombined in different ways to give rise to subtly different functions (Todd et al., 2001). The cellular and phenotypic functions cannot be deduced directly from protein structure but can be predicted based on biochemical functions. Thus, a number of residues in the protein may be associated with function, and this can be predicted and/or studied experimentally (Chapter 2). Functional assays can also be used to measure the functional output of proteins.

3.1.2 Formation of protein structure

A protein's native structure is defined as the overall 3D organisation or fold of its polypeptide chain. Protein folding is a spontaneous process which is governed by interatomic forces that act among and on its amino acid sequence. These interatomic forces include covalent disulphide bridges, electrostatic interactions, water shells and charged surface residues, hydrogen bonds, hydrophobic interaction and Van der Waals forces (temporary electric dipoles) (Branden & Tooze, 1999). For a protein, both the conformational folding reaction and the formation of disulphide bonds take place within the cellular endoplasmic reticulum, which is an oxidative environment (Narayan, 2012). This process is known as oxidative protein folding and is facilitated by a series of chaperones, enzymes and sensors that detect the presence of misfolded or unfolded proteins. Such enzymes include endoplasmic reticulum oxidoreduction-1 and protein disulphide isomerases. Protein folding may also begin on the ribosome whilst its polypeptide chain is still being synthesised.

For a protein, disulphide bonds, which form between the sulphur atoms of the thiol groups of two Cys residues, stabilise the structure and impose conformational rigidity. In general, proteins in the reduced environment of the cytoplasm contain free thiols (Trivedi et al., 2009). The length of a disulphide bond is typically 0.20 - 0.25 nm (2.0 - 2.5 Å) (Wiita et al., 2006).

For two charges, the magnitude of the electrostatic force is inversely proportional to the square of their separation distance, and the product of the charges. This is given by Coulomb's law. For electrostatic interactions, a type of ionic bond known as a salt bridge can form between spatially proximal pairs of oppositely charged residues. For proteins, a salt bridge can form between an anionic carboxylate (RCOO^-) group of either Asp or Glu and either the cationic ammonium (RNH_3^+) of Lys, the guanidinium ($\text{RNHC}(\text{NH}_2)_2^+$) of Arg, or nitrogen atom from His. To form the salt bridge, these charged-group atom centroids must lie within 0.4 nm (4 Å) of each other. Most stabilising salt bridges also contain at least one hydrogen bond between the side-chain charged-group atoms. For proteins, salt bridges are rarely found across segments joined by flexible linkers, which suggests that they constrain flexibility and motion. In contrast to salt bridges, nitrogen-oxygen (N-O) bridges only require at least one pair of side-chain functional-group nitrogen and oxygen atoms within a 0.4 nm (4 Å) distance of each other (Kumar &

[Nussinov, 2002a](#)). Longer-range ion pairs are weaker and often destabilising for proteins ([Kumar & Nussinov, 2002b](#)). Water can screen (or lessen) electrostatic interactions and this is known as its dielectric property ([Biedermannova & Schneider, 2015](#)).

Hydrogen bonds are formed between a hydrogen atom, which is covalently bonded to an electronegative atom (donor), and another electronegative atom (acceptor). For hydrogen bonds, the most frequently observed geometry corresponds to a distance of < 0.25 nm (2.5 Å) and a donor-hydrogen-acceptor angle of between 90° and 180° . For proteins, intramolecular hydrogen bonds are formed between main chain polar groups which lead to secondary structure elements ([Hubbard & Kamran Haider, 2010](#)). Hydrogen bonds are also formed between polar or charged residue side-chains and water, thus creating a protein surface network. For the surrounding hydration layer of a protein, each water molecule can simultaneously serve as an acceptor for up to two hydrogen bonds and a donor for an additional two hydrogen bonds ([Biedermannova & Schneider, 2015](#)).

For protein folding, both the hydrophobic interactions between hydrophobic side-chains and the interactions of polar and charged residues with water molecules are essential. Hydrophobic interactions occur over long ranges of 0-10 nm and decay exponentially with distance ([Israelachvili & Pashley, 1982](#)). Polar water molecules also interact directly with the protein via either the protein backbone or the side-chains of residues in the protein interior, or form clusters in hydrophobic cavities ([Ernst et al., 1995](#)). On average, for protein folding, 85% of non-polar side chains are buried ([Lesser & Rose, 1990](#)) and 1.1 hydrogen bonds per residue are formed ([Baker & Hubbard, 1984](#)). Mutations can affect the number of structural water molecules within the core and disrupt essential main-chain-water interaction networks, resulting in destabilisation of the protein ([Levy & Onuchic, 2004](#)).

Van der Waals or dispersion forces are weak interactions between temporary (induced) dipoles. These occur between the hydrophobic (non-polar) side chains of the residues Ala, Ile, Leu, Phe, Tyr, Val and Gly. For two non-bonding atoms, a mathematical model known as the Lennard-Jones potential describes the potential interaction energy as a function of the separation distance:

$$u_{L-J} = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (3.1)$$

By this, when the non-polar groups are at a distance, they do not interact, because they are not close enough to distort the other's electron cloud and produce dipoles. As they approach each other more closely, and attract each other (long-range), the potential energy decreases until a minimum is reached. If the atoms are forced any closer (short-range), there is a steep rise in energy from the electron clouds starting to overlap and repulsion results.

Interatomic forces are dependent on both the pH and temperature of the surrounding environment. Thus, for proteins, variations in pH can alter the ionisation states of amino acid side chains and lead to different electrostatic forces and the disruption of hydrogen bonds. In addition, chemical denaturants such as ionic surfactants or high concentrations of urea or guanidinium chloride cause loss of native protein conformations and lead to complete unfolding. Proteins generally aggregate after heat denaturation. At temperatures higher than 80 °C, the heat denaturation of proteins is usually irreversible. However, naturally occurring hyper-thermostable proteins that can withstand very high temperatures also exist, such as the hsCutA1 protein in humans. This may be due to an unusually high content of charged residues ([Matsuura et al., 2015](#)).

3.1.3 Protein stability

The stability of a protein is defined as the tendency to maintain its native (folded) structure. Anfinsen's classic experiment concluded from numerous denaturation-renaturation experiments that the native state of a protein corresponds to the global minimum of free energy and is reached in a pathway-independent manner. This was called the thermodynamic hypothesis of protein folding ([Anfinsen, 1973](#)). In general, thermodynamics describes the stability of molecules between different states by using the free energy quantity and does not depend on time. On the other hand, kinetics describes how fast a reaction will reach equilibrium and overcome the activation energy quantity. Despite a reaction being thermodynamically favourable, it may be kinetically unfavourable. For a chemical reaction, the Gibbs free energy (G) is the maximum energy that can be spent and its change provides information on the spontaneity. If the change in Gibbs free energy (ΔG) is negative, the reaction is thermodynamically favoured. For the

equilibrium between the native folded (N) and unfolded (U) states of a protein, the ratio of the folding (k_f) and unfolding (k_u) kinetic rate constants is known as the equilibrium constant K_{eq} . The change in Gibbs free energy for a population of molecules is calculated from the equilibrium constant by:

$$\Delta G = -RT \ln K_{eq} \quad (3.2)$$

where R ($\text{J K}^{-1}\text{mol}^{-1}$) is the gas constant and T (K) is temperature. Thus, protein stability is proportional to the Gibbs free energy change. The Gibbs free energy change for a process at constant pressure is:

$$\Delta G = \Delta H - T\Delta S \quad (3.3)$$

where ΔH and ΔS are the changes in enthalpy and entropy, respectively, and T is the temperature. For the folding of a globular (roughly spherical and mostly water-soluble) protein, the enthalpy (ΔH) is contributed by intra-molecular interactions whereas the configurational entropy (ΔS) is determined by the number of accessible configurations. For protein folding, a net decrease in the Gibbs free energy occurs due to both the internal covalent bonds stabilising the structure and the hydrophobic effect which leads to a stabilising decrease in entropy. These override the folding-unfavourable decrease in configurational entropy which favours the unfolded, more random state. At a constant (biological) temperature of $\sim 300\text{K}$, both the effective potential energy (ΔG) and the configurational entropy (ΔS) decrease as the native folded state is approached (Karplus, 2011). For globular proteins in physiological conditions, the Gibbs free energy change is typically small, of the order of 5-15 kcal/mol. Thus native proteins are only marginally stable entities.

The protein folding process can be described as a free energy surface (or landscape) which is a rugged funnel shape (Figure 3.1). By this, the denatured state of the protein populates a large ensemble of structures (high entropy) whereas the native state populates one or few conformational states (Bartlett & Radford, 2009). Here, ‘ensemble’ refers to a collection of structures that have been sampled. For the native state to correspond to multiple conformational states, the potential energy surface is characterised by a large number of thermally accessible minima close to that of the native structure (Elber & Karplus, 1987). For the native state, numerous folding pathways may

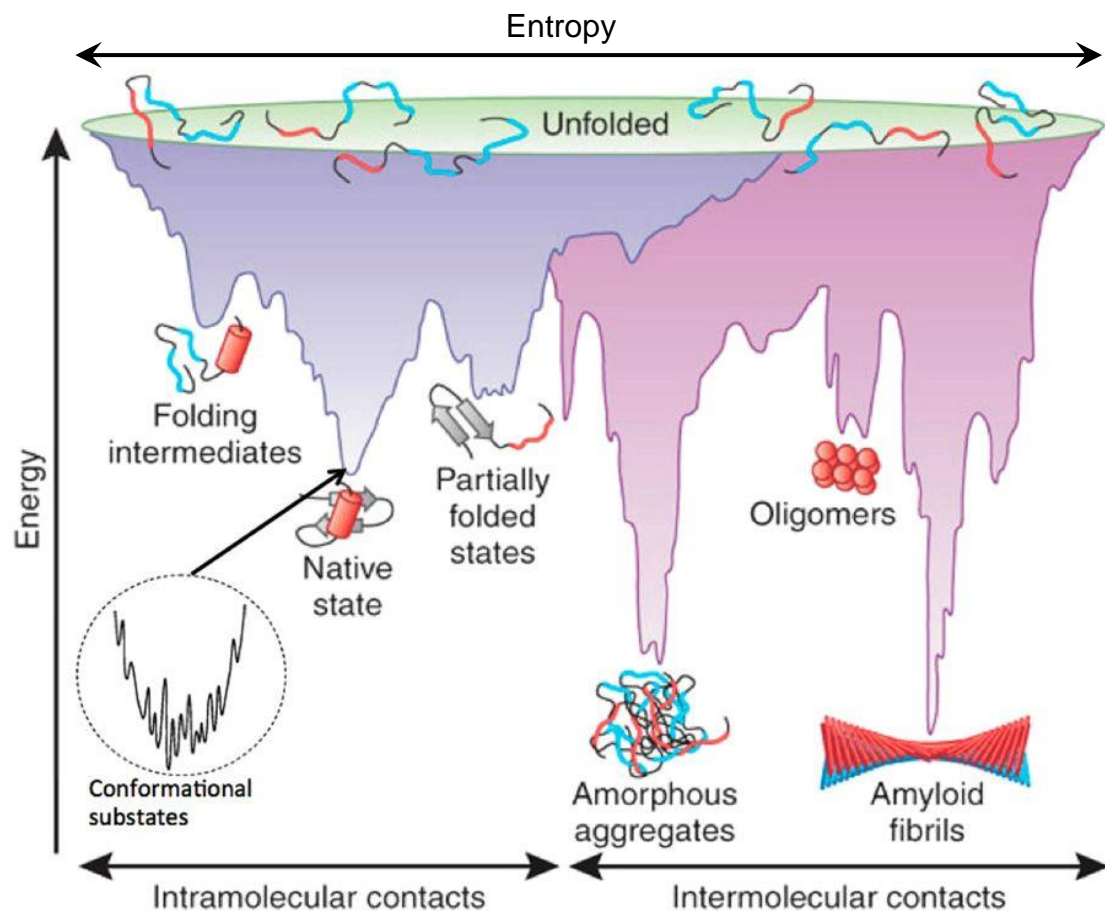


Figure 3.1 The folding funnel energy landscape for a globular protein. The top of the funnel (high potential energy) corresponds to unfolded proteins. As the protein folds, the entropy of the system, determined by the number of accessible configurations, is unfavourably decreased. However, a net decrease in potential energy (x-axis) occurs because the structure is stabilised by a combination of the formation of either internal interactions or intermolecular contacts and the hydrophobic effect. For native intramolecular contacts (x-axis, left), the native state has the lowest potential energy and may comprise a number of conformational substates (arrowed) around the local minima. For intermolecular contacts (x-axis, right), protein forms with the lowest potential energies include amorphous aggregates and amyloid fibrils. Adapted from (Raskatov & Teplow, 2017) and (Bartlett & Radford, 2009).

be present whereby multiple partially folded ensembles are populated *en route* (Onuchic & Wolynes, 2004). In addition, the protein may get trapped in local minima due to the formation of favourable but non-native (“wrong”) contacts, such as non-native hydrophobic contacts or hydrogen bonds which are favoured in terms of enthalpy. From the native to the unfolded state, the reversible thermal or melting transition of a protein can be determined by calorimetry methods. By this, the stabilising effect of disulphide bonds and other molecular interactions on proteins is manifested by an increase in the melting point of the protein.

In order for a protein to fold into the native (functional) state, a random search through all of the possible folding routes would take a very long time. This is because the length of a typical polypeptide chain introduces many degrees of freedom. For example, in order to estimate the time required for folding, the number of configurations of the polypeptide chain can be multiplied by the time required to find one configuration. For a 100 residue protein, this corresponds to 10^{70} configurations multiplied by 10^{-11} seconds which gives approximately 10^{52} years (Karplus, 1997). However, in nature, the speed of protein folding is rapid, being only a few seconds or less. This is known as Levinthal’s paradox (Levinthal, 1968). In order for a protein to achieve its native fold so rapidly, the native fold is thought to be favoured by local interactions of residues which limit the conformational space available for exploration. By this, protein folding is not an unbiased, random process. In contrast to Anfinsen, Levinthal postulated that a protein must follow a specific path that guides it to the native state under kinetic control, possibly to a local minimum. Thus, the native structure corresponds to a rapidly reachable minimum rather than the global one (Finkelstein, 2018). In order to resolve Levinthal’s paradox, a number of models have been developed including the nucleation-growth mechanism (Wetlaufer, 1973), the diffusion-collision model (Karplus & Weaver, 1994), the framework model (Kim & Baldwin, 1982) and the jigsaw-puzzle model (Harrison & Durbin, 1985). However, both thermodynamic and kinetic hypotheses are used for models of protein folding. Protein misfolding is seen in diseases such as Alzheimer’s disease, Creutzfeldt–Jakob disease, and bovine spongiform encephalopathy, and has been attributed to kinetic traps or folding to an alternate state of lower energy (Thomas et al., 1995). For rigid globular proteins, mutations associated with diseases of protein deposition have been shown to destabilise the native structure thereby increasing the concentration of partially folded or disordered conformers (Uversky & Dunker, 2010).

For a bead model of a polypeptide chain in two dimensions, fast folding was shown to occur if the stabilising interactions corresponded to those present in the native state, by the use of Monte Carlo (MC) simulations (Go & Abe, 1981). For the folding reaction of a protein, the first full analysis of the effective energy, entropy and free energy surface was also modelled by a MC simulation of a 27-bead heteropolymer with random interactions (Sali et al., 1994). This found that, for at least 30 out of 200 random sequences folded on a lattice, less time steps were required when compared to sampling all configurations.

Once folded, the structure of a protein can be described in terms of secondary structure elements and how they pack and connect together (topology) to form domains and tertiary and quaternary structures (Section 2.4). Proteins can have either single-domain or multi-domain modular architectures. For protein domains, the ‘Class, Architecture, Topology/fold, Homologous superfamily’ (CATH) database has classified 95 million protein domains into 6,119 superfamilies based on their evolution ancestries by using both structure and sequence (Sillitoe et al., 2015).

3.1.4 Protein dynamics

Proteins must be able to move in order to fulfil their biochemical functions in solution. By this, proteins are dynamic entities, and their motions are intrinsically encoded in the sequence of amino acids within the polypeptide chain (Frauenfelder et al., 1991; Karplus & Kuriyan, 2005). Proteins and other biomolecules are thus inherently flexible systems that display a broad range of dynamics which occur on time-scales of femtoseconds to seconds (Markwick et al., 2008). Protein dynamics range from small atomic fluctuations around an average structure to large-scale reorganisations and often contribute to conformational changes. Such conformational changes include structural fluctuations, allosteric changes, domain motions, local folding to unfolding transitions, secondary structure element transitions, disorder to order transitions and other sidechain and backbone changes (Ruvinsky et al., 2012). In a dynamic equilibrium, proteins rapidly interconvert between conformational states. Most eukaryotic proteins are ordered thus possess a 3D structure which is relatively stable with Ramachandran angles that vary slightly around equilibrium positions due to thermal energy. They may also have occasional co-operative conformational switches. However, more than a third of eukaryotic proteins have been shown to contain intrinsically disordered regions of over

30 residues in length (Ward et al., 2004). For an intrinsically disordered protein, the atom positions and Ramachandran angles vary significantly over time without specific equilibrium values and hence the protein lacks a stable 3D structure. Such intrinsically disordered proteins exist as dynamic ensembles and typically undergo non-cooperative conformational changes (Uversky & Dunker, 2010). For FH, 19 unstructured and flexible linkers of three to eight residues length are situated between the 20 globular SCR domains. These linkers are defined as intrinsically disordered regions. For a protein, by knowing all conformational sub-states and their associated populations under physiological conditions, protein function and biological processes can be better understood. For ordered proteins, the relative probabilities of the conformational states (thermodynamics) and the energy barriers between them (kinetics) can be defined as an energy landscape (Henzler-Wildman & Kern, 2007).

3.2 Methods for protein structure determination

For experimentally determined atomic-level 3D structures of biological macromolecules such as proteins, the Protein Data Bank (PDB) is the only global archive. The PDB is managed by the Worldwide Protein Data Bank organisation which includes three data centres in Europe, the United States and Japan (Rose et al., 2017). However, for proteins, the number of sequences deposited in the UniProtKB database (Bairoch et al., 2005) greatly exceeds that of structures in the PDB (Berman et al., 2000). By this, in 2015, there were approximately 90 million protein sequences compared to only about 100,000 protein structures. The reason behind this is that calculating the structure is much more demanding than determining the sequence, hence the importance of the work in this PhD thesis on determining the structure of FH. In order to elucidate the structure and/or dynamics of proteins, biophysical and structural biology techniques such as X-ray crystallography, small-angle scattering (SAS) of X-rays (SAXS) or neutrons, nuclear magnetic resonance (NMR) and cryo-electron microscopy (cryo-EM) can be employed. For protein sequences, the locations of structural domains can be predicted by using the CATH and Gene3D databases for bioinformatics-based analyses (Lees et al., 2012). By this, structure and function can be inferred from similar, well characterised proteins. For structures in the PDB in March 2017, 89% were determined by X-ray crystallography, 9% by NMR, 1% by cryo-EM and <1% by other techniques (Gore et al., 2017). For highly dynamic and flexible systems in solution, both NMR and SAS are considered the most appropriate tools. These types of systems are often problematic for EM and X-ray

crystallography methods. In terms of protein size, for NMR, structure determination is limited to moderately sized macromolecules (e.g. <50 kDa). In contrast, SAS does not have any size limitations. In terms of resolution, SAS is a low resolution technique, whereas NMR, cryo-EM and X-ray crystallography are higher resolution.

3.2.1 X-ray diffraction and scattering

For the diffraction technique of X-ray crystallography, the electromagnetic radiation is composed of X-rays (of approximately 1 Å wavelength) in order to probe interatomic distances. For the scattering technique of SAS, X-rays or neutrons (of approximately 1 Å or 10 Å wavelength, respectively) are used to probe the overall shape of particles in solution. Conventionally, X-rays can be generated by decelerating electrons whereby a target such as Cu is bombarded with an electron beam in order to produce intense, monochromatic peaks of specific wavelength. A monochromator is used to further define the X-ray wavelength. Often, SAS experiments are performed at synchrotrons.

During the X-ray scattering process, X-rays interact with electrons in the protein sample. By this, the electrons oscillate and become dipoles thereby emitting electromagnetic radiation of the frequency of the irradiated wavelength. From all of the oscillating electrons in the sample, an incident beam of X-rays thereby generates either scattered or diffracted X-rays, depending on the particle dynamics. Thus, for X-ray crystallography, the static crystal lattice arrangement of the particles produces a coherent diffraction pattern, whereas for SAS, the solution-based particles are tumbling in time which produces a less coherent scattering pattern. For the scattered X-rays, the intensity is proportional to the square of the charge/mass ratio of the particle. Particles with higher charge/mass ratios, such as electrons when compared to atomic nuclei or protons, are much more efficient at scattering. The relationship between scattering angle and inter-planar or inter-atomic spacing is given by Bragg's law:

$$n \lambda = 2 d \sin \theta \quad (3.4)$$

where λ is the radiation wavelength, d is the inter-planar spacing, θ is the angle between either the incident or diffracted ray and the relevant planes and n is the order of diffraction. The scattered waves interfere which produces distinct spots at specific angles. For the

diffraction space, there is an inverse relationship between the spacing in the object and the angle of diffraction which is usually called "reciprocal space". Peaks in the intensity of scattered radiation will occur when rays from successive planes interfere constructively according to Bragg's law.

3.2.2 Small angle scattering

For SAS, which was introduced in the 1930's by André Guinier, a distinct advantage is the speed of both data collection (ms; millisecond) on a modern synchrotron and sample characterization. In addition, the protein can be analysed in native conditions, including its flexibility and dynamics. For structure determination, several overall parameters can be extracted from the resulting radially averaged scattering pattern. However, the major limitation of SAS is that only low resolution models can be obtained and there will be ambiguity in the modelling (many models may fit the data). SAS can be combined with higher resolution methods for model validation and refinement by using hybrid modelling techniques. In addition, SAS-driven rigid body modelling routines can refine or determine the positions and orientations of protein domains or stretches of nucleic acids ([Mertens & Svergun, 2017](#)). For SAS, milligram (typically 1 – 2 mg) amounts of highly pure, monodisperse protein that remains soluble at high concentration are required. This ensures that there is no interaction between the particles and the particles are free to move.

For small angle neutron scattering, the excess scattering length density (contrast) is due to the nuclear scattering length density. Samples that highly absorb X-rays, such as solvents containing high salt, can be measured using small angle neutron scattering, and the samples will not suffer from radiation damage. For a protein, in order to study the structure and/or dynamics, hydrogen/deuterium (^2H or D) exchange can be used. Deuterium is a heavy isotope of hydrogen. By this, labile hydrogen atoms from either the OH-, NH-, and SH- groups of polar amino acid side chains or the NH-group of peptide bonds are replaced by deuterium atoms in the solvent ([Claesen & Burzykowski, 2017](#)). In small angle neutron scattering only and not SAXS, hydrogen/deuterium exchange leads to contrast variation neutron scattering. A disadvantage of small angle neutron scattering is the high incoherent scattering of hydrogen due to its negative coherent scattering length. This can often make the buffer subtraction difficult. In order to minimise this

incoherent scattering, deuterium, which has a positive scattering length, is used instead of hydrogen.

In SAXS, the sample is exposed to a collimated beam of incident radiation of a defined wavelength (typically 1 Å or 0.1 nm). Following Bragg's law (equation 3.4), at smaller angles of 0.1° to 0.5° used in SAXS, distances of approximately 100 Å can be detected. Thus, the smaller the angle, the greater the inter-atomic distance resolved. By this, the dominant scattering process is elastic, for which net energy transfer between the incident wave and the sample is not observed and the wavelength and energy of the scattered wave does not change. Other scattering phenomena such as Compton and Raman scattering of X-rays contributes to the background of the experiment. In SAXS, the scattering of X-rays by a solution of biomolecules is dependent on the number of particles, the size of the particles, and the contrast with the solvent. By this, the scattering is dependent on the concentration of the solute, the square of the volume of the individual particle and the square of the excess scattering length density ($\Delta\rho(r)$), respectively. The excess scattering length density refers to the difference in electron density ($\rho(r)$) between the solute and the solvent (contrast). For proteins, the electron density is approximately 410 e/nm³ which is only slightly greater than that of water at 334 e/nm³. For dilute aqueous solutions of proteins or other macromolecules, an isotropic scattering intensity results (Mertens & Svergun, 2017). This scattering depends on the modulus of the momentum transfer (or vector/direction), s (or Q):

$$s = 4\pi \sin(\theta)/\lambda \quad (3.5)$$

where 2θ is the angle between the incident and scattered beam. For s as a scattering vector:

$$\vec{s} = (\vec{k}_s - \vec{k}_i) \quad (3.6)$$

where k_s and k_i are the scattered and incident waves, respectively (Figure 3.2). For an atom, the scattering length density depends on the number of electrons, the distribution of the electrons and the atomic mass. Thus, the atomic scattering factor is a measurement of the scattering power of the atom and varies with X-ray wavelength. The atomic scattering factor is also the Fourier transform of the atom's spatial distribution or electron density. The Fourier transform is a mathematical transformation which inverts the units of the input variable, e.g. between time and frequency (1/time), or between space

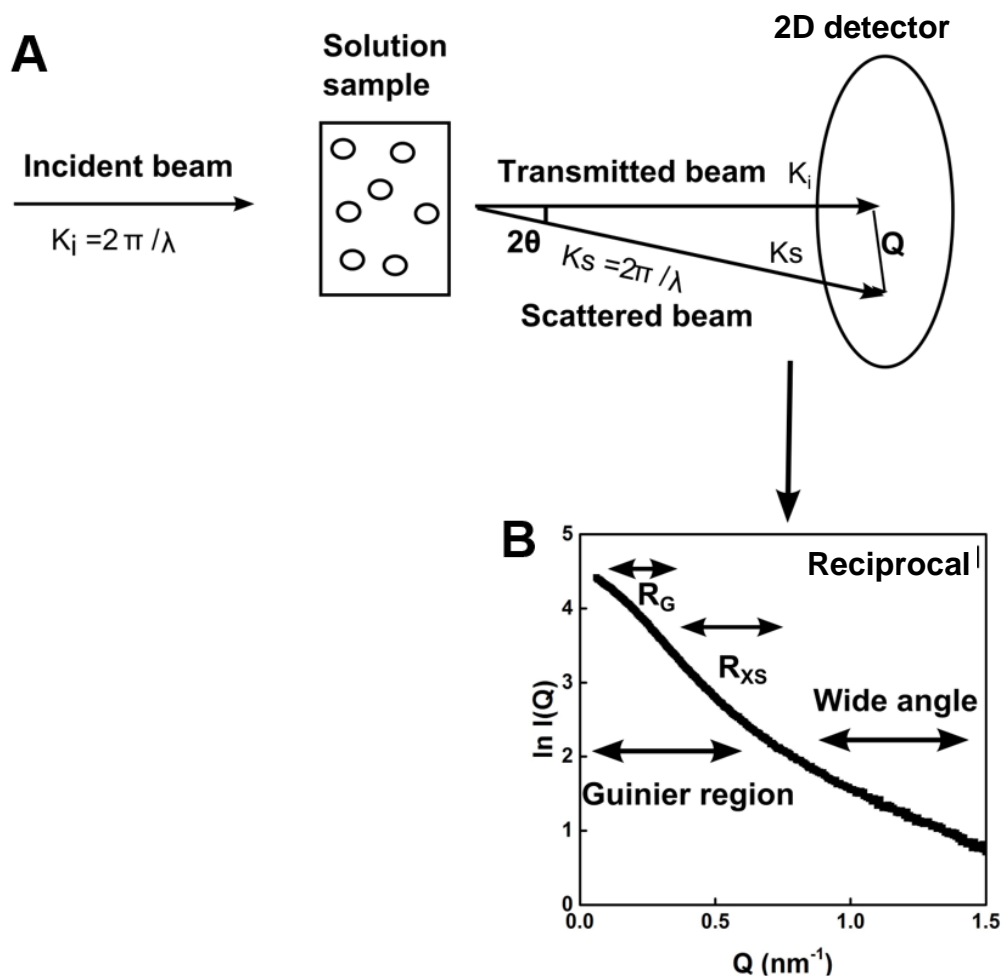


Figure 3.2 A schematic representation of a small angle scattering experiment (A) An incident beam of X-rays is scattered by the protein sample in solution. The 2D detector records the intensity $I(Q)$ of the scattered radiation and the scattering vector Q , which corresponds to the angle 2θ between the incident (transmitted) and scattered X-ray beams. **(B)** The scattering pattern is radially averaged to a 1 dimensional scattering curve that represents the reciprocal space. The R_G and R_{XS} parameters are calculated from specific Q ranges. Sourced from the PhD thesis of Dr Orla Dunne ([Dunne, 2015](#)).

(distance) and inverse-space (1/distance). The Fourier transform of the real space electron density profile probed by the X-rays is the observed scattering pattern (of constructive and destructive interference) in reciprocal-space. For a single scattering process, the scattering amplitudes of all atoms are summed up and the observed scattering intensity is given by an ensemble average of the intensity. The intensity at each Q is calculated by the Debye equation (Yang, 2014), which takes into account the differential orientations of the particles via rotational averaging in space:

$$I(Q) = \sum_p \sum_q f_p f_q \frac{\sin(rQ)}{rQ} \quad (3.7)$$

where f_p and f_q are the scattering lengths of the electrons at points p and q within the biomolecule, separated by a distance, r . For low concentrations of protein (approximately 1 mg/ml), the interference of X-rays from within molecules dominates the scattering profiles thereby reflecting the molecular structure. In contrast, higher concentrations of protein (above 1 mg/ml) leads to a significant amount of interference between X-rays scattered from different molecules which is known as inter-particle interference (Goldenberg & Argyle, 2014) and particularly shows at low angles ($Q < 1 \text{ nm}^{-1}$).

The scattered X-rays are detected by two-dimensional (2D) area detectors and radially averaged to one-dimensional (1D) SAXS profiles via integration for analysis. Experimentally, the detector can only measure an intensity, $I(Q)$, which is proportional to the squared amplitude of the scattered radiation. The resulting profile corresponds to the intensity, $I(Q)$, on the y-axis and the scattering angle, Q , on the x-axis. In order to account for variance in sampling distribution, multiple scattering curves are merged by averaging. After the solvent (buffer) scattering has been subtracted, the difference profiles are used for the direct extraction of important parameters for structural characterisation. These parameters provide information on the size, oligomeric state and overall shape of the molecules in solution. They include the molecular mass, radius of gyration (R_G), hydrated particle volume (V_p) and maximum particle diameter (D_{max}). For a protein, the R_G (\AA or nm) is defined as the root mean square distance from each atom to the protein's centre of mass. The centre of mass is also known as the centre of gravity. The R_G can be used as an indicator of compactness of the particle thereby the larger the R_G , the less compact the structure (Mertens & Svergun, 2017).

3.2.3 Guinier analyses for small angle scattering

The R_G can be determined at small angles for a monodisperse solution by using the Guinier approximation (Guinier et al., 1955) (valid when $Q \cdot R_G < 1.3$):

$$I(Q) \sim I_0 e^{\left(\frac{-Q^2 R_G^2}{3}\right)} \quad (3.8)$$

$$\ln I(Q) \sim \ln I_0 - \frac{Q^2 R_G^2}{3} \quad (3.9)$$

where $I(Q)$ is the scattering intensity, I_0 is the forward scattering intensity and Q (\AA^{-1} or nm^{-1}) is equal to s (equations 3.5 and 3.6). From the Guinier plot of $\ln I(Q)$ vs. Q^2 , the values of I_0 and R_G can be determined from the y-axis intercept and the slope of the linear region, respectively (equation 3.9) (Mertens & Svergun, 2017). For the molar mass of the protein, I_0 gives an independent estimation, being proportional to the number of electrons in the scattering particle. From the scattering curve, aggregation and inter-particle repulsion can also be identified via increased or decreased R_G values. For macromolecules which have an elongated shape, it is possible to calculate the mean cross-sectional radius of gyration (R_{XS}) from fitting a larger Q range (Figure 3.2) by using:

$$\ln[I(Q)Q] = \ln[I(Q)Q]_{Q \rightarrow 0} - \frac{R_{XS}^2 Q^2}{2} \quad (3.10)$$

The distance distribution function, $P(r)$, shows the frequency of pair-wise interatomic distances in the sample. By this, particle shape characteristics can be identified, e.g. globular particles which yield bell-shaped profiles with a maximum of approximately $D_{max}/2$ are spherical. In addition, multi-domain particles often yield $P(r)$ profiles with multiple shoulders that correspond to intra- and inter- subunit distances. Theoretically, a Fourier transformation of the scattering intensity (in reciprocal space) yields the distance distribution function (in real space), $P(r)$. However, in order to compute the $P(r)$, the indirect Fourier transformation is used because a direct Fourier transformation is not possible using only the finite number of points available from the scattering curve. This is because at low Q , the data points close to the beam stop will be missing and at high Q , the data points are limited to the edge of the detector. The indirect Fourier transformation method was first proposed by Glatter (Glatter, 1977). The method

parameterizes $P(r)$ in the range $[0, D_{max}]$, with D_{max} being a user defined variable, and determines the coefficients as a smooth function which provides the best fit to the experimental data. For this, automated programs such as GNOM ([Semenyuk & Svergun, 1991](#)) can be used:

$$P(r) = \frac{I}{2\pi^2} \int_0^\infty I(Q) Q r \sin(Qr) dQ \quad (3.11)$$

In addition, $I(0)$ and R_G can be directly computed from the $P(r)$, which is typically more accurate than using a Guinier analysis because it corresponds to real space ([Mertens & Svergun, 2017](#)).

3.2.4 Other methods

For X-ray crystallography, the diffraction of X-rays results in an electron density map which reflects the structural arrangement of atoms within the crystal. The signal to noise ratio is increased because a crystal arranges many molecules in the same orientation for the scattered waves to add up in phase. However, critical phase data is lost which prevents full retrieval of the real space 3D structure. This is known as the “phase problem”, and this can be overcome by methods such as molecular replacement. Overall, X-ray crystallography is a high resolution technique that has helped to elucidate the majority of protein structures in the PDB at the atomic level. However, a disadvantage of X-ray crystallography is that flexible and integral membrane proteins are very difficult to crystallise and purify. *In vivo*, many proteins are also glycosylated by flexible carbohydrate structures. Furthermore, proteins may not be able to be produced in sufficient quantities for crystallisation, and the method of crystallisation can introduce artefacts whereby the protein is forced into an unnatural frozen state, sometimes with the use of artificial ligands. The crystallisation of multiple functionally relevant states of proteins may also be difficult ([Nogales & Scheres, 2015](#)). The nature of the crystals such as size of the unit cell and lack of order may also make structural determination difficult.

In cryo-EM, proteins can be observed directly in multiple conformations in their native environment at near atomic resolution ([Murata & Wolf, 2018](#)). A high-energy beam of electrons (wavelength ~ 1 nm or 10 \AA), which are accelerated by an electrostatic potential, interacts with a sample and the elastic interactions are detected. By this, the

density of the sample results from the projected Coulomb potential or the 2D projection along the beam direction (Carroni & Saibil, 2016). For cryo-EM, biological specimens are solidified without ice formation by suspension in solid water via the use of liquid nitrogen. When kept at below -160°C , the protein sample is stable in the high vacuum and also more resistant to electron beam damage. Proteins are very weak at scattering electrons thus in order to see the (unstained) protein, a phase contrast effect is produced from using defocus and spherical aberration with a solvent such as dilute salt solution (Saibil, 1996). For the 3D protein, numerous 2D projections are obtained which are of random unknown directions and very noisy. In order to reconstruct the 3D volume, complicated post-processing involving classifying, aligning and averaging by using statistics is required. Compared to SAXS which is fast and easy but of lower resolution, a cryo-EM experiment yields a higher resolution 3D density map (Kim et al., 2017). Depending on instrumental limitations, cryo-EM can provide structural information in a resolution range of $\sim 3 \text{ \AA}$ to $\sim 30 \text{ \AA}$ (Hong-Wei & Jia-Wei, 2017). For heterogeneous samples, the division of particle images into more homogenous subsets can be carried out by using analyses such as eigenimage sorting (Villarreal & Stewart, 2014).

For NMR, the resonating frequencies or chemical shifts of nuclei (such as ^1H , ^{13}C and ^{15}N for proteins) are dependent on the local magnetic field thus reflect the local molecular environment. By this, the quantum mechanical properties of atoms, such as the nuclear spin, are exploited. NMR can be used to probe average inter-atomic distances of either 6 \AA or less based on the Nuclear Overhauser Effect (Wüthrich, 1986) or longer-range based on paramagnetic relaxation techniques (Iwahara & Clore, 2006). By combining these experimental NMR data with theoretical models or a library of short fragments of known 3D structure, a 3D protein structure can be determined (Markwick et al., 2008).

3.3 Atomistic molecular modelling

In order to elucidate the native structure of a protein, data collected from biophysical experiments are fitted to molecular models generated by computational methods. For molecular models in biology, the level of detail or resolution varies from atomistic-scale to overall shape. The process of moving up the scale at the cost of finer detail is known as coarse graining. In order to generate protein models, either *ab initio* or *de novo* ('from scratch') modelling, or homology or template-based modelling can be

used ([Sali & Blundell, 1993](#)). Both modelling methods require the sequence of the protein, but the latter method also requires a structure for a homologous protein to act as a structural template.

3.3.1 Homology modelling

For homology modelling, the 3D structure of a given protein sequence (target) is predicted based primarily on its alignment to one or more proteins of known structure (templates) that are typically homologous (> 30%). By this, the folds are assigned, the target and templates are then aligned and finally the model is built and evaluated ([Webb & Sali, 2016](#)). For retrieval of homologous sequences with 3D structures, the multiple sequence alignment algorithm BLAST can be used to search the PDB database and the template structural PDB file can be downloaded. Multiple templates can also be used for the homology modelling. The quality of the resulting structure is highly dependent on the sequence alignment. At low sequence identity (30 - 35%) the chance of errors increases for both the alignment and the model. Thus, the alignment should be visually inspected to ensure that the most conserved residues are well aligned. Generally, the loop regions have low sequence similarity and can be highly variable in terms of length and dynamics. In order to optimise the sequence alignment, the CLUSTAL OMEGA ([Sievers et al., 2011](#)) alignment program is used followed by manual checks. For the homology modelling, the MODELLER software ([Webb & Sali, 2016](#)) is one of the most widely used and can be accessed both as a webserver and as a Python library which can be customised. MODELLER also features a variable loop optimisation protocol and models can be selected based on their comparison to an atomic distance-dependent reference statistical potential (free energy function) known as the discrete optimised protein energy score. The most recent statistical potentials are determined by using statistics on the frequencies of pairwise residue contacts for proteins in the PDB and are based on the Boltzmann energy distribution. For the discrete optimised protein energy score, the sizes of the reference native proteins are also taken into account ([Shen & Sali, 2006](#)). In order to improve protein characterisation, the output homology model can be further refined, studied and verified by using a combination of molecular dynamics and experimental data.

3.3.2 *Ab initio* modelling

For *ab initio* modelling, the native structure of the protein is computationally predicted from its amino acid sequence by using physicochemical principles. By this, *ab initio* modelling also helps address how and why a protein adopts a specific structure out of many possibilities. However, due to both physical and evolutionary constraints, the number of possible folds is limited, and predicted to be in the range of 1000 – 10,000 folds (Grant et al., 2004). Thus, in order to quicken the *ab initio* method, possible folds are identified based on a comparison of the target sequence with a set of related protein structures, in a process known as threading. For *ab initio* modelling, the output depends on three main factors. Firstly, in order to identify the native structure from all possible structures, an accurate energy function is required to find the most thermodynamically stable state. For energy functions, either physics-based or knowledge-based functions are used which depend on whether they make use of existing structures in the PDB. An example of a knowledge-based function is the rule that proline residues are uncommon in α -helices. Secondly, the low-energy states must be identified quickly via an efficient conformational search method, and thirdly, near-native models need to be selected from a pool of non-native structures by using a certain strategy (Rigden, 2009) which may include statistical clustering and/or free-energy based methods. *Ab initio*-based programs for modelling include the ROSETTA (physics and knowledge-based) (Simons et al., 1997) (Das et al., 2007) and I-TASSER (knowledge-based) (Roy et al., 2010) algorithms.

In order to check the quality of protein structural models, a number of computational tools exist. For checking the residue torsion angles, Ramachandran plot analyses identify any outlier residues with very unusual backbone conformations (Ramachandran et al., 1963). For a high quality dataset, the number of outlier residues is less than 0.5% (Gore et al., 2017). Energy minimisation is effective in relaxing any high energy, strained inter-atomic interactions (Section 3.3.4).

3.3.3 Molecular dynamics

For a protein structure, molecular dynamics (MD) simulation can be used for both *ab initio* modelling, thus simulating folding directly from sequence (Section 3.3.2), and studying the physical movements of the folded state over time. However, normally, *ab initio* refers to quantum mechanics based modelling, and although quantum mechanics

based MD simulations are possible, they would be prohibitively expensive for molecules the size of a protein. Hence, usually classical MD simulations, based on Newtonian mechanics, are employed for large biomolecules. For a MD simulation of a system of interacting particles (atoms), the trajectories of molecules, thus the updated atomic positions, are determined by numerically solving Newton's equations of motion (classical mechanics) and by using a force field. For the force field, for each atom, a mathematical formulation of the potential energy contains terms which are associated with bond lengths, angles, torsion angles, van der Waals and electrostatic interactions. By this, force fields are essentially huge databases of molecular properties which describe the detailed mathematical functions that account for the forces between atoms. They consist of both bonded components, which includes stretching, bending, torsion and improper interactions, and non-bonded components, which includes both electrostatic and Van der Waals. Stretching and bending interactions can be described by a harmonic potential. The force field treats atoms as point particles interacting through a defined potential form. The end result is a very large potential energy expression. Current computational resources do not allow the use of physics-based quantum mechanics. However, the parameters which govern inter-atomic interactions are obtained from comparisons of the force field with both experimental and quantum mechanical data ([Hagler et al., 1974](#); [Weiner et al., 1984](#)). For all-atom physics-based force fields, well known examples include Assisted Model Building with Energy Refinement (*AMBER*) ([Case et al., 2005](#)), *CHARMM* ([Brooks et al., 2009](#)), and *GROMOS* ([Reif et al., 2013](#)). Between these force fields, the major differences correspond to both the selection of atom types and the interaction parameters.

MD simulations are made possible by the multitude of algorithms and computer programs which are closely coupled to advances in both parallel computing and system hardware. Limitations to MD include very long simulations and very large systems, however present simulation times are close to biologically relevant timescales. MD simulations are highly sensitive to the parameters set by the user and can be performed in a large set of different conditions such as vacuum, implicit solvent, explicit water or saline solution of physiological salt strength and lipid membranes ([Hospital et al., 2015](#)).

For the modelling of the environment surrounding proteins, which consists of water and ions, simulations are either explicit or implicit. For explicit models, both water molecules and ions are included, whereas for implicit models, only the mean force exerted by the external environment or media onto the protein is approximated. By this, implicit

models contribute either none or only a few degrees of freedom to the simulation thus speed up the computation time. However, important detailed features such as hydrogen bond fluctuations, conformational changes which reorient water dipoles and bridging water molecules are neglected ([Kleijnung & Fraternali, 2014](#)).

By using MD for studying folded proteins, the dynamic evolution of the system resulting from atomic interactions in physiological conditions (such as solvent and temperature) over time is computationally simulated. By using MD, the energy surface is explored in order to find alternative configurations. The lower energy states are more probable than the higher energy ones, according to the Boltzmann distribution, but both may appear in the ensemble via thermal fluctuations. The Boltzmann distribution gives the probability of the system being in a certain state as a function of that state's energy. MD is thus an important tool for studying the conformational ensembles of proteins. However, the true global minimum may not be found. For exploring the conformational space, the traditional approach of cumulating static experimental structures ("experimental ensembles") allows only a partial view of flexibility and may be biased ([Hospital et al., 2015](#)). However, MD can be compared with experimental data, such as strength of binding characteristics, for verification purposes. MD simulation was first developed in the late 1950s within the field of theoretical physics ([Alder & Wainwright, 1959](#)). The first biological system to be studied by MD was the folded globular protein bovine pancreatic trypsin inhibitor, for which the equations of motion were solved for the atoms with an empirical potential energy function ([McCammon et al., 1977](#)). In terms of timescales, the time-step (speed limit) of MD is of the order of femtosecond. If atomic vibration is slow, the time-step can be a few femtoseconds. In increasing order of time, typical biomolecular movements include bond stretching (femtosecond to picosecond), elastic vibrations, rotations of surface sidechains and hinge bending (picosecond to nanosecond), and allosteric transitions, local denaturation and rotation of buried side chains (microsecond to second). For MD simulation of large biomolecular systems, Nanoscale Molecular Dynamics (NAMD) is a parallel MD code which was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign ([Phillips et al., 2005](#)). NAMD produces output files that contain information about the co-ordinates, energies and velocities of each atom. The output DCD file contains the trajectory of the system. For running simulations in NAMD, a configuration file sets the parameters required such as how long-range electrostatic interactions are treated and the

dielectric variable for H₂O. For force field parameter files, NAMD2 supports *CHARMM*. For NAMD, a protein structure file (PSF) contains information about the molecular structure such as the bonds, angles, dihedrals and impropers. PSF files can be built using topology files or generated by using the *CHARMM-GUI* online (Jo et al., 2008). By this, the structural PDB files are parameterised and compatible for *CHARMM*.

Many proteins are glycosylated, such as FH with eight bi-antennary di-sialylated glycans which are Asn linked (Fenaille et al., 2007). In SAXS, glycan positions on proteins can be identified however their visualisation has been limited due to the similar scattering contrast of carbohydrate relative to protein (Hammel et al., 2002). Despite this, a reconstruction of the overall shape of the glycoprotein showed that the glycans appeared to either contribute bulk volume or just a bulge near the known glycan attachment site (Guttman et al., 2013). In order to prepare protein structural models with glycan chains, the program GlycanReader (Jo et al., 2011) can generate the *CHARMM* force field and PSF inputs required for MD and/or energy minimisation of the glycans in NAMD. Following this, energy-minimised glycans can be positioned onto structures by using PyMol scripting, and re-run through GlycanReader to generate a new set of input files. For the resulting PDB files, any nomenclature or numbering changes required for other modelling programs can be carried out by using bash scripting.

3.3.4 Energy minimisation

Initial atomic coordinates of protein structures may have unphysical contacts which lead to high energies and forces. Thus, by energy minimising the whole system, the temperature is decreased to 0°C which decreases the kinetic energy. By this, the overall energy is decreased and any high energy interactions are relaxed. In terms of the energy landscape, a nearby local minimum is found, and further equilibration to nearby states is possible via low energy barriers. In order to access these other intermediate states in a timely manner in a MD simulation, steering forces such as random thermal fluctuations at physiological temperatures are required. NAMD can also be used for energy minimisation, before running an MD simulation.

3.3.5 Monte Carlo methods

MD simulations can take a very long time and lots of computing power. In order to sample the configuration space or energy landscape more quickly, MC allows intermediate states to be sampled without solving the equations of motion for each step. By this, MC can be described as the “highlights” of an MD simulation and provides efficient sampling without considering time. As described by Boltzmann’s Ergodic hypothesis, the time-average of a quantity calculated by MD is equivalent to the ensemble-average of the same quantity calculated by MC. By this, points in configuration space are generated with a relative (to zero energy) probability that is proportional to the Boltzmann factor.

For a given structure in an energy minimum, random perturbation using MC generates a trial structure. In order to generate the trial structure, the user defines both the flexible regions that will move and sets the degree to which those flexible regions will move. For proteins, the ϕ and/or ψ torsion angles around the peptide bond (protein backbone) are varied, in keeping with the physiochemical restraints of peptide bonds in the Ramachandran plot ([Ramachandran et al., 1963](#)). By this, in the MC simulations, these bonds can act as either hinges or they are twisted (torsion angle). For the peptide bond, the omega (ω) angle is always set to 180° .

In order to determine systematically the acceptance of the trial structure over the first (starting) structure, the energy difference between them can be calculated by using the Metropolis algorithm ([Rigden, 2009](#)). By this, if the change of potential energy (ΔE) associated with the displacement of atoms is less than zero (i.e. new structure is of a lower energy state), the new co-ordinates are accepted. If ΔE is equal to or greater than zero and the Boltzmann factor (probability of a state of energy E , relative to the state of zero energy) for the structure is less than a randomly generated number, the new coordinates are accepted. By this, some higher energy structures (which are less probable according to the Boltzmann distribution) are randomly accepted and may contribute to a high energy barrier being overcome to get to another local minima. This reflects the physiological random thermal fluctuations. The longer the MC simulation is run, the more likely this will happen. Otherwise, the new coordinates are discarded and the old coordinates are kept to start the process again. Structures with steric clashes are of high energy and will be discarded. MC simulations of the system are prone to becoming stuck in meta-stable

states. In order to prevent this from distorting the distribution of sampled states, the energy barriers can also be overcome by sampling states in a wide range of energies by changing the temperature.

3.3.6 The SASSIE workflow

For SAXS, in order to identify conformational ensembles of structures, the scattering curves can be compared to theoretical curves calculated for physically realistic atomistic-scale molecular models. Such models will have been obtained from rigorous energy minimisation, MD and/or MC simulations at physiological conditions. By this, although SAXS is typically limited to shape parameters due to both being a low resolution technique and loss of information from orientation averaging, atomistic-scale molecular models that produce very similar theoretical SAXS scattering patterns can be identified. Such models are then not only consistent with the scattering data but also with some of the known physical chemistry of the system.

Traditional structural analyses for SAS make assumptions about the class of structure, such as radius and density/composition for a sphere, in order to fit the experimental data. By this, a large number of form factors for different shapes have been derived. However, these are restricted to shape classes with simple symmetry whereas biomaterials such as proteins have little or no symmetry. Thus, a method whereby experimental data is fitted to shape envelopes from spherical harmonics or assemblies of small spheres was developed by EMBL in the ATSAS suite of programs ([Petoukhov et al., 2012](#)). The first atomistic scale modelling was based on crystal structures but constrained by scattering curves and showed that only a limited number of antibody fragment conformations fitted the experimental scattering data. In order to automate this process, the SCT and SCTPL software tools ([Perkins et al., 2011](#); [Perkins et al., 2008](#)) were developed by using a commercial MD package and applied to human antibody structures. This allowed the first atomistic scattering structure to be deposited in the PDB. However, these methods are not suitable for analysing the flexibility of proteins within a conformational ensemble. More recently, in order to facilitate the analysis of SAXS data by using atomistic-scale biomolecular modelling, the Collaborative Computational Project for SAS (CCP-SAS) has developed software, deployment infrastructure and a workflow framework. SASSIE-web combines all of the necessary tools such as preparing protein structures, carrying out simulations, calculating theoretical scattering data and comparing the results with the

experimental data (Figure 3.3). A key feature of SASSIE is the dihedral angle MC simulation methods via Markov sampling of protein backbone torsion angles (Curtis et al., 2012). For MC, typically, around 10,000 to 50,000 structures are generated for analyses. In order to avoid steric overlap, a distance cut-off (typically 0.3 nm) is specified. Because MC is a coarse-grained method, further refinement of the structure is usually necessary. For MD and energy minimisation simulations, NAMD is used as the simulation engine. For the solvent, an implicit model with an efficient solvent-accessible surface area (Ferrara et al., 2002) that is applicable to proteins is used. For the models, in order to calculate the theoretical scattering curve from the atomic positions, the Debye equation is used. This requires atomistic pairwise distances to be calculated for every atom which is computationally demanding and time consuming. Both coarse graining of the original atomic structures and binning algorithms accelerate this process by reducing the number of scattering centres and number of distances to be processed, respectively. In SASSIE, the SasCalc module calculates scattering profiles from a structure file by using an exact all-atom expression for the scattering intensity. By this, the orientations of the Q vectors are taken from a quasi-uniform spherical grid generated by the golden ratio (Watson & Curtis, 2013).

In order to identify the best-fitting structures, the theoretical scattering curves are compared to the experimental curves by using the R -factor metric:

$$R = \sum_{Q_i} \frac{\|I_{expt}(Q_i) - I_{theor}(Q_i)\|}{\|I_{expt}(Q_i)\|} \times 100 \quad (3.12)$$

where $I_{Expt}(Q)$ and $I_{Theor}(Q)$ are each the intensity at the Q value for the experimental and the theoretical scattering curves, respectively. By the use of data interpolation, the two scattering curves have equal Q spacing. The R -factor is a percentage, and models with lower R -factors show the best fits thus inferred to represent the average solution structure (Wright & Perkins, 2015).

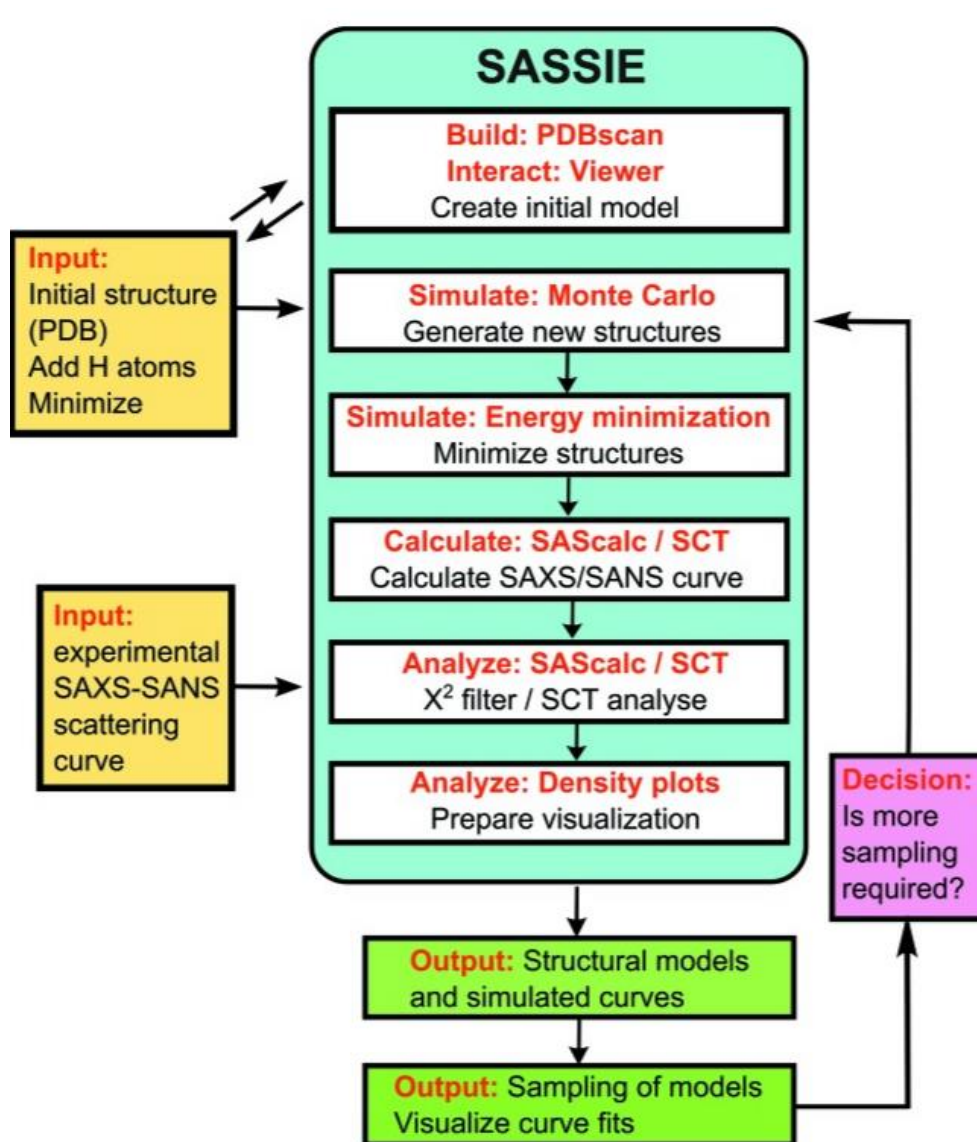


Figure 3.3 The SASSIE workflow. A schematic representation of the inputs, modules and outputs for a typical modelling project in SASSIE. The decision (pink box) is determined by the amount of structures that well fit the experimental data. Sourced from (Perkins et al., 2016).

3.3.7 Methods for analysing conformational ensembles

After filtering out theoretical scattering curves that are not well-fitted to the experimental data, by applying filters such as an R -factor $<5\%$ and R_G within 5% of the experimental R_G , a heterogeneous ensemble of different best-fit structures may result. These best-fitting structures can be seen as corresponding to the co-existence of multiple conformational states for a flexible protein in solution. These structures will be encoded in a trajectory file, most commonly in either DCD or PDB format. In order to characterise the conformational ensemble of best-fit structures in the trajectory, each major conformational state must be identified and quantified. The program Visual Molecular Dynamics (VMD) ([Humphrey et al., 1996](#)) can be used to visualise and analyse trajectory files. By using the Tcl scripting facility in VMD, custom code can be developed and applied to the trajectory file in order to gain information on the structure and dynamics of each conformational state. By this, structural features of the protein such as inter-domain movements and organisations can be identified geometrically by calculating the distances between certain atoms and/or the centre of mass. For SAXS, the R_G parameter cannot discriminate between two molecules that despite being similar in overall shape, thus appearing symmetrical, also show significant conformational differences. For example, for a long flexible protein such as complement factor H in solution, conformational states that correspond to either the N-terminal or C-terminal bending inwards are observed ([Nan et al., 2010](#); [Okemefuna et al., 2009](#)).

In order to calculate, visualise and compare inter-atomic and/or centre of mass distance distributions, the statistical program R provides a number of packages ([R Core Team, 2013](#)). By this, sampling distribution analyses can further be used to verify that the MC set up did not unexpectedly bias the simulation. If unexpected bias is present, a population of conformational states may have been ignored, and the simulation may need repeating with a different set of parameters to increase the sampling. Examples of such parameters include the length of the simulation, the maximum torsion angles and the definition of the flexible regions. For measuring the dynamics of the protein, thus the movement over time, the root mean square deviation of each atom when compared to the initial structure can probe protein flexibility and equilibration. In addition, the root mean square fluctuation, in which the reference is the mean structure over the trajectory, can be used to find conserved positions of C- α atoms in the models ([Benson & Daggett, 2012](#)).

In order to gain a more detailed analysis of the different conformational states sampled in the trajectory, the statistical technique principal component analysis (PCA) can be employed ([Hui et al., 2015](#)). For PCA, a covariance matrix is used to characterise the variance between frames in the trajectory. During PCA, multi-dimensional patterns in the dataset are found and subsequently compressed by reducing the number of dimensions without losing much information. In order to do this, eigenvectors and eigenvalues are used. The eigenvector with the highest eigenvalue is the principle component (PC) of the dataset and represents the most significant relationship between the datasets (dimensions). Thus, for protein structures, the first principal component corresponds to the greatest variance between the structures. Overall, the greater the difference in displacement between two frames, the further away they will be from each other along the principal component line. This allows the best-fit structures to be clustered by their atomic co-ordinates, and the most variant regions to be mapped onto the structures. PCA has been used for the characterisation of a number of proteins ([Gendoo & Harrison, 2012](#); [Yang et al., 2009](#)) and can be applied to datasets by using the Bio3D package in R ([Grant et al., 2006](#)). When combined with visualisation of the structural ensemble by using density plots, PCA can be used to identify the major structural conformations within ensembles.

An alternative method for finding a combination of structures that fits the experimental scattering curve is the minimal ensemble search method. By this, the smallest possible number of structures that recapitulate the experimental scattering data is attempted to be found ([Pelikan et al., 2009](#)). In order to compute the theoretical multi-conformational scattering from a minimal ensemble, the individual scattering patterns from the conformers are averaged. This is based on the ensemble optimization method which uses a genetic algorithm to select randomly generated models from a pool ([Bernado et al., 2007](#)).

3.3.8 Application to the complement proteins

For complement, a diverse set of proteins with different functions are required ([Chapter 1](#)). For the proteins associated with aHUS and C3G, which are mostly complement AP but also thrombosis-related, many different protein domains are involved ([Table 3.1](#)). The complement proteins FB, FH, the FHRs and MCP are composed of SCR domains, also known as sushi domains or complement control protein modules. SCR

Table 3.1 Complement and related protein domains

Protein ^a	Uniprot ID	Domain type	Amount
C3	P01024	Anaphylatoxin-like (ANA)	1
C3	P01024	Anchor	1
C3	P01024	alpha N-terminal loop (a'NT)	1
C3	P01024	C345C	1
C3	P01024	C1r/C1s-Uegf-Bmp1 alpha (CUBa)	2
C3	P01024	Macroglobulin (MG)	9
C3	P01024	Thioester domain (TED)	1
DGKE	P52429	C1	2
DGKE	P52429	Kinase accessory domain (DAGKa)	1
DGKE	P52429	Kinase catalytic domain (DAGKc)	1
FB	P00751	Linker helix aL (aL)	1
FB	P00751	Short complement regulator (SCR)	3
FB	P00751	Serine protease	1
FB	P00751	von Willebrand factor type A (VWFA)	1
FH	P08603	Short complement regulator (SCR)	20
FHR1	Q03591	Short complement regulator (SCR)	5
FHR2	P36980	Short complement regulator (SCR)	4
FHR3	Q02985	Short complement regulator (SCR)	5
FHR4	Q92496	Short complement regulator (SCR)	9
FHR5	Q9BXR6	Short complement regulator (SCR)	9
FI	P05156	FI membrane attack complex (FIMAC)	1
FI	P05156	Low density lipoprotein-receptor class A (LDLR)	2
FI	P05156	Serine protease	1
FI	P05156	Scavenger receptor cysteine-rich (SRCR)	1
MCP	P15529	Cytoplasmic tail (CT)	1
MCP	P15529	Linker and Ser/Thr rich region	1
MCP	P15529	Short complement regulator (SCR)	4
MCP	P15529	Transmembrane (TM)	1
Plasminogen	P00747	Kringle	5
Plasminogen	P00747	Plasminogen-apple-nemotode (PAN)	1
Plasminogen	P00747	Peptidase serine 1 (Peptidase S1)	1
Properdin	P27918	Thrombospondin type-1 (TSP type-1)	7
THBD	P07204	C-type lectin	1
THBD	P07204	Epidermal growth factor-like (EGF-like)	6
THBD	P07204	Integrin binding domain	1

^aProtein name abbreviations include: DGKE, diacylglycerol kinase epsilon; FB, factor B; FH, factor H; FHR, factor-H related protein; FI, factor I; MCP, membrane cofactor protein; THBD, thrombomodulin.

domains are evolutionary conserved and based on a β -sandwich topology. A β -sandwich is composed of two opposing antiparallel β -sheets, and is one of the most common structural motifs in the proteins of multicellular organisms. It was first identified in the domains of Ig structures such as antibody and hence called the Ig-fold. For the SCR domain, its β -sandwich arrangement is typically composed of six β -strands which envelope a hydrophobic core. For each of SCR-11 and SCR-12 of FH, their six and five β -strand arrangements, respectively, are shown in [Figure 3.4](#). For the SCR domains of FH, FHR1-5 and MCP, four conserved Cys residues co-ordinate two sets of disulphide bonds for stability ([green, Figure 3.4](#)). In addition, each SCR domain has one conserved Trp residue ([purple, Figure 3.4](#)). The SCR domain is associated with both heparin and complement C3b binding functions.

For the complement AP and related proteins that are involved in aHUS and C3G, many have been structurally characterised to a certain extent by using a range of biophysical methods ([Table 3.2](#)). However, to date, full-length FH has proven too large, flexible and glycosylated for high resolution structural determination by X-ray crystallography or NMR. Preliminary molecular modelling of SAXS curves revealed that full-length FH in solution has a folded-back SCR structure that is affected by ionic strength ([Aslam & Perkins, 2001](#); [Nan et al., 2010](#); [Okemefuna et al., 2009](#)). The structures of each of DGKE, the FHRs 2 to 5 and plasminogen have not yet been solved.

For this PhD thesis, in order to examine the overall domain arrangement of FH and how it is affected by the AMD-risk Tyr402His polymorphism and to better understand FH complement binding, the solution structures for both the His402 and Tyr402 FH allotypes were studied ([Chapter 5](#)). The resulting new models of FH in solution were used for my analyses in [Chapter 6](#) which aimed to identify whether the structural location of a variant in FH complexes may predict the occurrence of aHUS or C3G in patients. Furthermore, my new FH models can also be deposited in the Database of Complement Gene Variants ([Chapter 4](#)) for better structural-based analyses of novel complement gene variants.

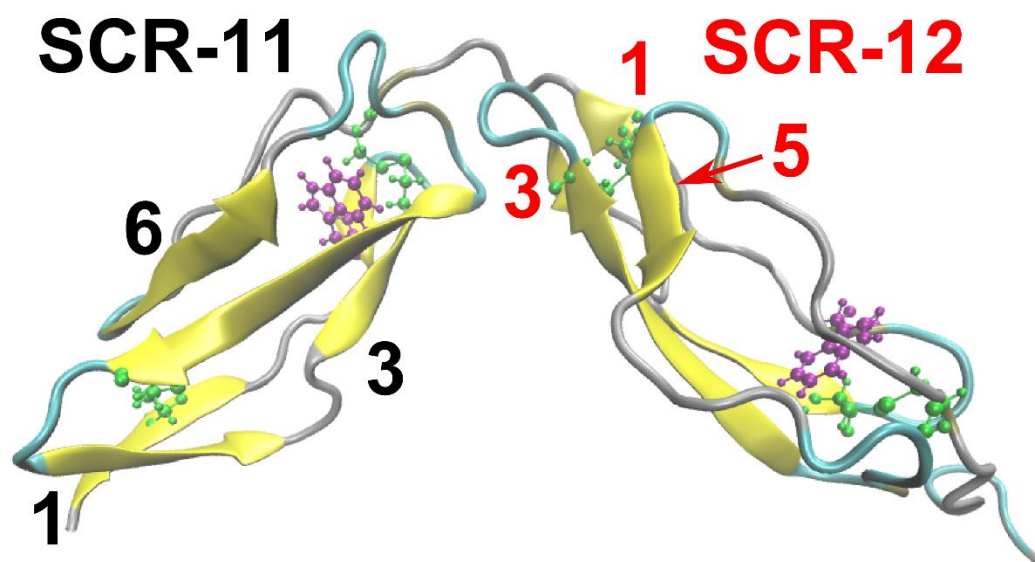


Figure 3.4 Structural arrangement of two short complement regulator domains of complement factor H. For each of the factor H short complement regulator (SCR) 11 and 12 domains, the structural arrangement of the model with Protein Data Bank code 4B2S is shown in cartoon style. The β -sheet structures are shown in yellow and numbered in black and red text, respectively. The green and purple ball and stick (CPK) style atomic representations show the four conserved Cys (two disulphide-bonded pairs) and one conserved Trp residue per domain. The cyan and grey coloured regions represent the turn and coil secondary structures, respectively.

Table 3.2 Representative molecular structures for the complement proteins

Protein ^a	Method ^b	Code ^c	Description	Resolution (Å)	Reference
C3	X-ray crystallography	2A73	Full length C3	3.3	(Janssen et al., 2005)
C3	X-ray crystallography	2I07	Full length C3b	4	(Janssen et al., 2005)
C3	X-ray crystallography	5FO8	Complement C3b in complex with MCP SCR-1/4	2.4	(Forneris et al., 2016)
C3	X-ray crystallography	5FO7	Complement C3b	2.8	(Forneris et al., 2016)
C3	X-ray crystallography	5FO9	Complement C3b in complex with CR1 SCR-15/17	3.3	(Forneris et al., 2016)
FB	X-ray crystallography	1RRK	Bb segment of FB	2	(Ponnuraj et al., 2004)
FB	X-ray crystallography	2OK5	Full length FB	2.3	(Milder et al., 2007)
FH	SAXS and AUC	3GAV	Full length FH	-	(Okemefuna et al., 2009)
FH	SAXS and AUC	3N0J	Full length FH	-	(Nan et al., 2010)
FHR1	X-ray crystallography	3ZD2	The two N-terminal domains of FHR1	1.99	(Goicoechea de Jorge et al., 2013)
FHR1	X-ray crystallography	4MUC	The fourth and fifth C-terminal domains of FHR1	2.9	(Bhattacharjee et al., 2015)
FI	X-ray crystallography	2XRC	Full length FI	2.6	(Roversi et al., 2011)
MCP	X-ray crystallography	3O8E	Extracellular region of CD46 in complex with human adenovirus type 11 fiber knob	2.8	(Persson et al., 2010)
Properdin	SAXS	1W0R	Full length properdin	-	(Sun et al., 2004)
Properdin	SAXS	1W0S	Full length properdin	-	(Sun et al., 2004)
THBD	X-ray crystallography	1DX5	THBD-thrombin complex	2.3	(Fuentes-Prior et al., 2000)
THBD	X-ray crystallography	3GIS	Na-free thrombin in complex with THBD	2.4	(Adams et al., 2009)

^a Protein name abbreviations include: FB, factor B; FH, factor H; FHR1, factor-H related protein 1; FI, factor I; MCP, membrane cofactor protein; THBD, thrombomodulin.

^b Method abbreviations include: SAXS, small angle x-ray scattering; AUC, analytical ultra-centrifugation.

^c Protein Data Bank code

Chapter Four

**Statistical validation of rare complement
variants provides insights on the
molecular basis of atypical haemolytic
uraemic syndrome and C3
glomerulopathy**

4.1 Summary

Atypical haemolytic uraemic syndrome (aHUS) and C3 glomerulopathy (C3G) are associated with dysregulation and over-activation of the complement alternative pathway. Typically, gene analysis for aHUS and C3G is undertaken in small patient numbers, yet it is unclear which genes most frequently predispose to aHUS or C3G. Accordingly, a six-centre analysis of 610 rare genetic variants in 13 mostly complement genes (*CFH*, *CFI*, *CD46*, *C3*, *CFB*, *CFHR1*, *CFHR3*, *CFHR4*, *CFHR5*, *CFP*, *PLG*, *DGKE*, and *THBD*) from >3500 patients with aHUS and C3G was performed. There were 371 novel rare variants for aHUS and 82 for C3G. A new interactive Database of Complement Gene Variants was used to extract allele frequency data for these 13 genes using the Exome Aggregation Consortium (ExAC) server as the reference genome. For aHUS, significantly more protein-altering rare variation was found in five genes *CFH*, *CFI*, *CD46*, *C3* and *DGKE* than in ExAC (allele frequency <0.01%), thus correlating these with aHUS. For C3G, an association was only found for rare variants in *C3* and the N-terminal C3b-binding or C-terminal non-surface-associated regions of *CFH*. In conclusion, the rare variant analyses showed non-random distributions over the affected proteins, and different distributions were observed between aHUS and C3G that clarify their phenotypes.

4.2 Introduction

aHUS and C3G are two severe ultra-rare renal diseases that involve dysregulation of the AP in the complement system of innate immunity. In healthy individuals, the AP eliminates unwanted pathogens without compromising host cells due to the balance between AP activator and regulatory proteins. aHUS and C3G both feature host cell attack by the AP leading to end stage renal failure. However, aHUS is a thrombotic microangiopathy with an acute presentation, whereas C3G is not a thrombotic microangiopathy, but is characterised by an abundance of C3 deposition in the renal glomeruli with a mostly chronic presentation, with some exceptions (Goodship et al., 2017).

Rare genetic abnormalities in AP and thrombosis-related genes are present in ~60% of aHUS cases (Bu et al., 2012; Fakhouri et al., 2017; Warwicker et al., 1998). These mostly drive AP dysregulation at the endothelial cell surface (Goodship et al., 2017). The penetrance of predisposing RVs in aHUS is ~50% and is determined by the *CFH* and *CD46* (MCP) haplotype, additional RVs, and a trigger (Bresin et al., 2013; Esparza-Gordillo et al., 2005; Noris & Remuzzi, 2015; Sansbury et al., 2014). As reported in the 2014 aHUS database (Rodriguez et al., 2014), most genetic aHUS cases are heterozygous (Nester et al., 2015) and are attributed to the genes *CFH* (25-30% of cases), followed by *CD46* (8-10%), *C3* and *CFI* (4-8% each) and *CFB* (1-4%) (Goicoechea de Jorge et al., 2007; Noris & Remuzzi, 2015). The C-terminal SCR-19/20 domains of FH are a well-known RV hotspot for aHUS, this being attributed to its functional binding sites for C3b, C3d and heparin (Noris & Remuzzi, 2017). In distinction to other genes, most RVs in *CD46* lead to a quantitative decrease in the protein product and approximately a quarter are homozygous (Nester et al., 2015). Rare copy number variation leading to large genomic rearrangements in the *CFHR*-*CFH* region, such as the *CFH/CFHR1* and *CFH/CFHR3* hybrid genes, are risk factors for aHUS (Eyler et al., 2013; Maga et al., 2011; Valoti et al., 2015; Venables et al., 2006). RVs in the non-complement gene *THBD* account for 3-4% of genetic aHUS (Noris & Remuzzi, 2015), although no *THBD* variants were detected in the French aHUS cohort (Fremeaux-Bacchi et al., 2013). RVs in the non-complement gene *DGKE* account for ~27% of aHUS presenting under the age of one year and <4% under two years, respectively (Lemaire et al., 2013; Sanchez Chinchilla et al., 2014).

In C3G, complement gene RVs have been identified in ~20% of sporadic C3G cases (Bu et al., 2016; Iatropoulos et al., 2016; Noris & Remuzzi, 2015; Servais et al., 2012). Familial C3G is most often linked to highly penetrant heterozygous copy number variation in the *CFHR1-5* genes, such as *CFHR5* nephropathy, as well as homozygous *CFH* deficiency and heterozygous GoF mutation in *C3* (Athanasίου et al., 2011; Ault et al., 1997; Chauvet et al., 2016; Chen et al., 2014; Gale et al., 2010; Levy et al., 1986; Martinez-Barricarte et al., 2010; Tortajada et al., 2013). These frequently affect AP regulation in the fluid phase, with some exceptions (Noris & Remuzzi, 2017). As in aHUS, unaffected carriers of these genetic abnormalities are seen, indicating that the genetic variant only predisposes for the manifestation of C3G (Thomas et al., 2014). aHUS and C3G both also involve anti-CFH autoantibodies, known as acquired factors (5-13% of cases) (Durey et al., 2016; Noris & Remuzzi, 2015).

Genetic sequencing and multiple ligation-dependent probe assessment screening panels for aHUS and C3G typically include up to 10 complement (*CFH*, *CFHR1*, *CFHR3*, *CFHR4*, *CFHR5*, *CFI*, *C3*, *CD46*, *CFB* and complement factor properdin (*CFP*; *FP*) (Kouser et al., 2013)), two coagulation (*THBD* and *PLG* (Bu et al., 2014)) and one non-complement (*DGKE*) genes (Goodship et al., 2017). RVs in six genes (*CFH*, *CFI*, *C3*, *CD46*, *CFB* and *DGKE*) are associated with aHUS, while RVs in *CFH* and *C3* associate with C3G. These associations result from studies of many aHUS patients, and rather fewer C3G patients by linkage and familial segregation, case-control cohorts and functional studies. However, the associations of RVs in the four genes *CFHR5*, *CFP*, *PLG* and *THBD* with aHUS, and all 13 genes with C3G are less well defined (Goodship et al., 2017). In addition, an increasing number of variants of unknown significance for aHUS and C3G are being identified amongst these 13 genes. These require rapid pathogenicity evaluations for clinical interpretation, e.g. almost a third of *CFH* variants have limited functional characterisation (Merinero et al., 2017).

In order to clarify differences in the genetic and molecular basis of aHUS and C3G, I analysed the statistical AFs of RVs in multiple patient cohorts for comparison with reference datasets, and developed a new Database of Complement Gene Variants (<http://www.complement-db.org>). The AF provides the frequency of the variant in a given population. Variant pathogenicity can be identified by comparisons with genomic reference datasets such as the ExAC (Bennett et al., 2017). The database also provides structural biology and evolutionary tools to predict the effect of RVs on these proteins,

and present functional data from the literature and ClinVar. The database was used to analyse 610 rare genetic variants from >3500 patients in six renal centres, this being the largest dataset known to date for aHUS and C3G with 371 aHUS and 82 C3G novel RVs. Following comparisons with 60,706 genomic reference sequences from ExAC ([Lek et al., 2016](#)) and 6,500 from the EVS, I confirmed the associations of six genes (above) with aHUS, and found that three genes *CFHR5*, *PLG* and *THBD* were not associated with aHUS. The statistical comparisons also confirmed that *CFH* and *C3* were associated with C3G, and suggested the involvement of *CFB* and *THBD* with C3G. My results explain how changes in the same proteins result in the different pathologies observed with aHUS and C3G. Through the use of AFs and burden testing in my new database, this will inform patient management by enabling clinical immunologists to interpret new variants in terms of their associations with aHUS and C3G.

4.3 Methods

4.3.1 Data collection

The aHUS and C3G phenotype and variant data was sourced from six centres (unpublished and published data) and literature searches. aHUS was diagnosed by the presence of one or more episodes of microangiopathic haemolytic anaemia and thrombocytopenia defined on the basis of hematocrit <30%, hemoglobin <10mg/dl, serum lactate dehydrogenase >460U/L, undetectable haptoglobin, fragmented erythrocytes in the peripheral blood smear, and platelet count <150,000/ μ l, associated with acute renal failure, together with a negative Coombs test, ADAMTS13 activity >10% and negative Shiga toxin ([Goodship et al., 2017](#)). C3G was diagnosed by the presence of C3 deposits by immunofluorescence in the absence, or comparatively reduced presence of immunoglobulins. The identification of dense deposits within the glomerular basement membrane by electron microscopy lead to the further classification of the C3G as dense deposit disease ([Goodship et al., 2017](#)). The database included variant data in 13 genes (*C3*, *CD46*, *CFB*, *CFH*, *CFHR1*, *CFHR3*, *CFHR4*, *CFHR5*, *CFI*, *CFP*, *DGKE*, *PLG*, *THBD*) ([Table 4.1](#)). The estimated variant AFs in the aHUS and C3G datasets were based on data from the six centres only. Data from the three reference genome projects, namely ExAC (Version 0.3) ([Lek et al., 2016](#)), EVS (NHLBI GO Exome Sequencing Project, Seattle, WA (<http://evs.gs.washington.edu/EVS>), and 1000GP ([Genomes Project et al., 2015](#)), were downloaded and used as surrogate control reference datasets for all

Table 4.1 Summary of mRNA and protein identifiers for the 13 genes

Gene	Protein	RefSeq (NCBI) ^a		Ensembl ^a	
		mRNA	Protein	Transcript ID	Protein ID
<i>CFH</i>	FH	NM_000186.3	NP_000177.2	ENST00000367429	ENSP00000356399
<i>CFI</i>	FI	NM_000204.3	NP_000195.2	ENST00000394634	ENSP00000378130
<i>C3</i>	C3	NM_000064.3	NP_000055.2	ENST00000245907	ENSP00000245907
<i>CD46</i>	MCP	NM_002389.4	NP_002380.3	ENST00000358170	ENSP00000350893
<i>CFB</i>	FB	NM_001710.5	NP_001701.2	ENST00000425368	ENSP00000416561
<i>CFHR1</i>	FHR1	NM_002113.2	NP_002104.2	ENST00000320493	ENSP00000314299
<i>CFHR3</i>	FHR3	NM_021023.5	NP_066303.2	ENST00000367425	ENSP00000356395
<i>CFHR4</i>	FHR4	NM_001201550.2	NP_001188479.1	ENST00000367416	ENSP00000356386
<i>CFHR5</i>	FHR5	NM_030787.3	NP_110414.1	ENST00000256785	ENSP00000256785
<i>THBD</i>	THBD	NM_000361.2	NP_000352.1	ENST00000377103	ENSP00000366307
<i>CFP</i>	FP	NM_001145252.1	NP_001138724.1	ENST00000396992	ENSP00000380189
<i>DGKE</i>	DGKE	NM_003647.2	NP_003638.1	ENST00000284061	ENSP00000284061
<i>PLG</i>	Plasminogen	NM_000301	NP_000292	ENST00000308192	ENSP00000308938

^a The databases of annotated genomic, transcript and protein reference sequences are found at <https://www.ncbi.nlm.nih.gov/refseq/> (RefSeq) and <http://www.ensembl.org/index.html> (Ensembl).

genes apart from *CFHRI*, *CFHR3* and *CFHR4* that were involved in copy number variation only. These latter three genes were excluded because, at the time of writing, no copy number variation data were available from ExAC, EVS or 1000GP. These three reference datasets do not contain genomes from patients with rare renal disease phenotypes that include aHUS or C3G, however ExAC and EVS contain genomes from myocardial infarction studies. While this may affect the analyses of thrombosis-related genes such as *THBD*, a recent study found that the ExAC was not enriched in pathogenic RVs for these diseases (Song et al., 2016). In ExAC, only variants with “PASS” filter status were included in my analyses. The corresponding AF of each variant in ExAC, EVS and 1000GP was queried by using Human Genome Variation Society nucleotide level nomenclature (c.) and National Center for Biotechnology Information (NCBI) gene accession number. Published experimental data on each variant were sourced from literature searches using PubMed. Variants were described using DNA and protein level terms for both Human Genome Variation Society and legacy nomenclature.

4.3.2 Data cleansing, duplications and maintenance

Patient duplicate tests were carried out within and between datasets from the six renal centres, firstly by identifying potential duplicates using the patient’s variant profile, disease, gender and year of birth. Potential duplicates were investigated by requesting the full date of birth details from each renal centre and deleting as necessary. For maintenance of the database in the long term, each collaborating group will provide an annual data update as a spreadsheet that will be automatically uploaded to the database with automated duplication and error checks that are programmed into the MySQL backend. As an open resource, other clinical centres are invited to upload formatted data updates through the curator. To prevent against SQL injection attacks and related software vulnerabilities, the database inputs were sanitised and tested using the penetration testing tool SQLMAP.

4.3.3 Web-database development

The Database of Complement Gene Variants (<http://www.complement-db.org>) was constructed using a MySQL platform and its user interface was developed using a combination of PHP, JavaScript, jQuery, CSS and HTML. The design is based on the European Association for Haemophilia and Allied Disorders coagulation Factor IX web-

database (<http://www.factorix.org/>) (Rallapalli et al., 2013). The database is mounted on a Linux server within the Information Services Division at University College London.

4.3.4 Data retrieval

Variant data were retrieved from the Database of Complement Gene Variants using simple or advanced search tools. The user first defined the data source and gene, and then the other search options, including protein domain, location, and variant type and effect. The advanced search tool featured a customisable default ExAC AF cut-off value of 1%, which was used to ensure that any variant with an ExAC AF >1% was not retrieved. Variants were mapped to their genomic location using the reference human genomes GRCh37 and GRCh38. The protein and transcript locations were identified using RefSeq and Ensembl (Table 4.1). The database provided protein structural views of each missense variant using the JSmol Java applet (<https://sourceforge.net/projects/jsmol/>). These views facilitated a structural understanding of the variant using the webpage ‘structure and function’ tab. For each disease dataset, the statistics page showed (i) the distribution of variants in genes by protein domain, exon/intron location, and variant type and effect; (ii) the RV burden per gene, and (iii) the number of RVs per case. The AF webpage summarised the number of variants in each gene using their reference genome AF, with links to their variant database entries, for each disease dataset. The world map webpage showed the number of laboratory-sourced cases that have citizenship in each country, where data were available. The new variants webpage predicted the effects of any theoretical missense variants using AF, structural and functional analyses. The structures webpage listed all known protein structural models with links to their PDB code. The amino acid alignments webpage showed the multiple sequence alignment of up to ten vertebrate species for each protein depending on sequence availability, this being calculated using the UniProt database and the Probabilistic Alignment Kit program (PRANK) (Loytynoja & Goldman, 2005).

4.3.5 Rare variant burden

The RV burden per gene was computed for the aHUS, C3G, ExAC and EVS datasets. The RV burden was defined as the proportion of screened cases with an identified RV per gene. For aHUS and C3G, the burden was calculated by dividing the number of cases with an identified RV by the total number of cases screened per gene.

For ExAC and EVS, this was calculated for each gene by dividing the total mean adjusted allele count, this being determined from the allele count minus the number of homozygous cases, by the total mean adjusted number of subjects screened (Walsh et al., 2016). An AF cut-off value of 0.01% was used for all RV burden calculations, which is advised for a Mendelian disease (Kobayashi et al., 2017). The RV burden was calculated for all protein-altering RV only; truncating (nonsense, frameshift and splice) and non-truncating (missense and in-frame). RVs classified as ‘benign’ and ‘likely benign’ (see next section) were not filtered out because these data were unavailable for the ExAC and EVS datasets.

4.3.6 Rare variant assessment

Each non-large genomic rearrangement RV was classified as ‘pathogenic’, ‘likely pathogenic’, ‘uncertain significance’, ‘benign’ or ‘likely benign’ using categorisation guidelines (Goodship et al., 2017) that followed the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (Richards et al., 2015). The following categories were also added to cover all RVs in the datasets:

1. MAF <0.1%; likely benign by clinical testing (Illumina; ClinVar); *in silico* analyses predict: tolerated, neutral and benign. Categorised as ‘likely benign’ for all genes except for *C3* and *CFB* which were ‘uncertain significance’ instead.
2. MAF <0.1%; likely benign by clinical testing (Illumina; ClinVar); *in silico* analyses predict: uncertain deleterious effects. Categorised as ‘uncertain significance’.
3. MAF <0.1%, predicted as loss-of-function (includes nonsense, frameshift and splice acceptor/donor variants) in *C3* or *CFB*. Categorised as ‘uncertain significance’. *C3* or *CFB* loss-of-function is unlikely to over-activate complement and lead to aHUS or C3G. Since their effects on complement have not been experimentally proven, it was safer to categorise these as ‘uncertain significance’.
4. MAF <0.1%; synonymous change; *in silico* analyses predict: tolerated, neutral and benign. Categorised as ‘likely benign’.
5. 0.1% < MAF <1%; no functional data; in *C3* or *CFB*. Categorised as ‘likely benign’.

Loss-of-function RVs in all genes, except in those only involved in large genomic rearrangement variants or in the complement activators *C3* and *CFB*, were classified as ‘likely pathogenic’ for aHUS and C3G, unless functional data reported otherwise

(Goodship et al., 2017). The Ensembl Variant Effect Predictor tool (McLaren et al., 2016) was used for PolyPhen-2 (Adzhubei et al., 2010) and SIFT (Kumar et al., 2009) analyses and the results were used for combinatorial variant analyses. These classifications were made available for each variant identified in the aHUS and C3G datasets.

4.3.7 Spatial distribution of missense rare variants in the proteins

For each protein domain, the AFs of missense RVs in the aHUS and C3G datasets were summed, and then divided by the proportion of protein residues in the corresponding domain, in order to identify mutational hotspots. Missense RVs that were categorised as ‘benign’ or ‘likely benign’ were excluded from these analyses. For FH, FI, C3, MCP and FB, where there were enough missense RVs to identify mutational hotspots, these were represented graphically as bar charts. For FH, FI, C3, MCP and FB, each unique missense RV was mapped onto the structural model for visualisation, but noting that these did not take the aHUS or C3G AF into account. The structural models for FH (PDB code 3N0J (Okemefuna et al., 2009)), C3 (PDB code 2A73 (Janssen et al., 2005)), FI (PDB code 2XRC (Roversi et al., 2011)), MCP (PDB code 3O8E (Persson et al., 2010)) and FB (PDB code 2OK5 (Milder et al., 2007)) were sourced from the PDB. An alternative model for FH in which the C-terminal domains were extended and not the N-terminal domains was also used but not shown here (PDB code 3GAV) (Rodriguez et al., 2014). The R statistical package was used for statistical analyses and artwork (<http://www.R-project.org/>). PyMol was used for protein structural visualisation and artwork (Schrodinger, 2015).

4.3.8 Statistical analyses

The categorical variables shown in Tables 4.2 and 4.6, Appendix II, III and IV, and Figures 4.4 and 4.5 were examined using the two-tailed Chi-squared test (χ^2) with Yates’ correction with a 0.05 significance level that was Bonferroni-corrected where applicable. For the common variants analyses in Table 4.2, the 0.05 significance level was Bonferroni-corrected by dividing by the 14 variants, to give 0.0036. For the patient gender analyses in Table 4.6, the 0.05 significance level was Bonferroni-corrected by dividing by 12 genes to give 0.0042 for aHUS, and by 11 genes to give 0.0045 for C3G. The ‘ALL’ genes category for both aHUS and C3G were each subjected to a 0.05 significance level (no Bonferroni adjustment needed). For the RV burden analyses presented in Figure 4.4 and Appendix II and III, the 0.05 significance level was

Bonferroni-corrected by dividing by the 9 genes to give 0.0056. For the statistical analyses of *CFH* in [Appendix IV and Figure 4.5](#), a 0.05 significance level was used. The categorical variables shown in [Figure 4.2B and Table 4.3](#) were examined using the two-tailed Fisher's exact test with a significance level of 0.05.

For each row of one independent t-test (protein-altering only) in [Appendix II and III](#), I undertook a power calculation using the program PS Power and Sample Size Calculations (Version 3.0) ([Dupont & Plummer, 1990](#)). My calculations used expected differences of: $\geq 5\%$ for *CFH*, *C3* and *CD46*, $\geq 2.6\%$ for *DGKE* in the aHUS and EVS groups, and $\geq 2.5\%$ for *DGKE* in the other three groups, and $\geq 2\%$ for all other genes, and took into account the unequal sizes of the experimental (aHUS or C3G) and control (ExAC or EVS) groups. If the true difference between the experimental and control groups was as expected, I was able to reject the null hypothesis that their frequencies were equal with a certain probability (power). The Type I error probability associated with all of these tests was 0.0056. The power was at least 80% in all tests apart from the following cases in [Appendix II](#), with ExAC as the reference dataset, for aHUS: *DGKE* (74%), and for C3G: *DGKE* (55%) and *PLG* (32%), and also in [Appendix III](#), with EVS as the reference dataset, for aHUS: *DGKE* (67%), and for C3G: *DGKE* (68%), *PLG* (10%), *THBD* (38%) and *CFB* (41%).

4.3.9 Allele frequency (AF) analyses

A two-tailed Chi-squared test (χ^2) with Yates' correction was used to assess the difference in protein-altering RV AF (ExAC AF < 0.1%) between the aHUS/C3G and reference datasets. Protein-altering variants included non-truncating and truncating variants only ([Walsh et al., 2016](#)). For each RV, the 0.05 significance level was Bonferroni-adjusted by dividing by the number of RVs identified in its gene for aHUS and ExAC (*C3*: 485, *CD46*: 191, *CFB*: 400, *CFH*: 469, *CFHR5*: 262, *CFI*: 215, *DGKE*: 193, *PLG*: 263, *THBD*: 145), and for C3G and ExAC (*C3*: 452, *CD46*: 130, *CFB*: 391, *CFH*: 334, *CFHR5*: 261, *CFI*: 173, *DGKE*: 176, *PLG*: 263, *THBD*: 143) and rounding to 4 decimal places to give significance levels of 0.0001, 0.0002, 0.0003 or 0.0004 accordingly. The AF was not available in the aHUS dataset for the two *CFP* RVs so *CFP* was not analysed. *CFHR1*, *CFHR3* and *CFHR4* were not analysed because large genomic rearrangements were not included in ExAC at the time of writing. Related individuals were taken out the analyses by including familial alleles only once (twice for homozygous

cases). This procedure refers to the AF assessments that were displayed on the web database, and were also used to verify that none of the RVs were significantly more common in ExAC than in aHUS or C3G.

4.4 Results

4.4.1 Genetic variants in aHUS and C3G

In order to perform the comparisons between aHUS and C3G using AF analyses, six Reference Centres from the United Kingdom, France, Italy, Spain, Holland and the United States of America provided phenotype and variant data sets for 3128 and 443 patients from National and International Registries. The total of disease variants was 543 for aHUS and 111 for C3G in 13 genes ([Figure 4.1](#)). The aHUS and C3G variants were analysed in terms of their reference AFs for these variants found in three reference datasets (1000GP, EVS and ExAC). First, 10, 10 and 8 common variants with AF > 1% were filtered out ([red in Figure 4.1](#)). The RVs that were retained were those with an AF of <1% in the reference datasets ([light blue, green, dark blue, and pink, Figure 4.1](#)), or were rare large genomic rearrangements not covered by the reference datasets. By this filtering, a total of 610 retained RVs (96-97% of the total) were identified in 13 genes ([Table 4.1](#)) for aHUS (542 variants) and C3G (110 variants), with 42 being shared by aHUS and C3G, including the large genomic rearrangements.

For this PhD thesis chapter, only RVs were analysed for aHUS and C3G. The combined presence of multiple variants that are common in the general population can modify the risk for aHUS ([Ermini et al., 2012](#)), such as the SNP c.1204**T**>C p.Tyr402His (rs1061170) in the CFH-H3 haplotype ([Pickering et al., 2007](#); [Rodriguez de Cordoba & Goicoechea de Jorge, 2008](#)). The CFH-H3 aHUS-risk haplotype also contains the SNPs -331**C**>**T** (rs3753394), c.184**G**>A Val62Ile (rs800292), c.2016**A**>**G** p.Gln672Gln (rs3753396), IVS15 -543**G**>A intron 15 (rs1410996) and c.2808**G**>**T** p.Glu936Asp (rs1065489), where the at-risk alleles are defined in bold. In contrast, the presence of one RV can predispose for genetic aHUS and C3G. Thus, the analyses of multiple common variants per case for aHUS and C3G are more complex and were beyond the scope of this PhD thesis. For the aHUS and C3G datasets, the remaining 3-4% of common variants with AF >1% in at least one of the reference datasets comprised 14 and four in aHUS and C3G respectively ([Table 4.2](#)). These 14 and four variants were only just over the 1% AF

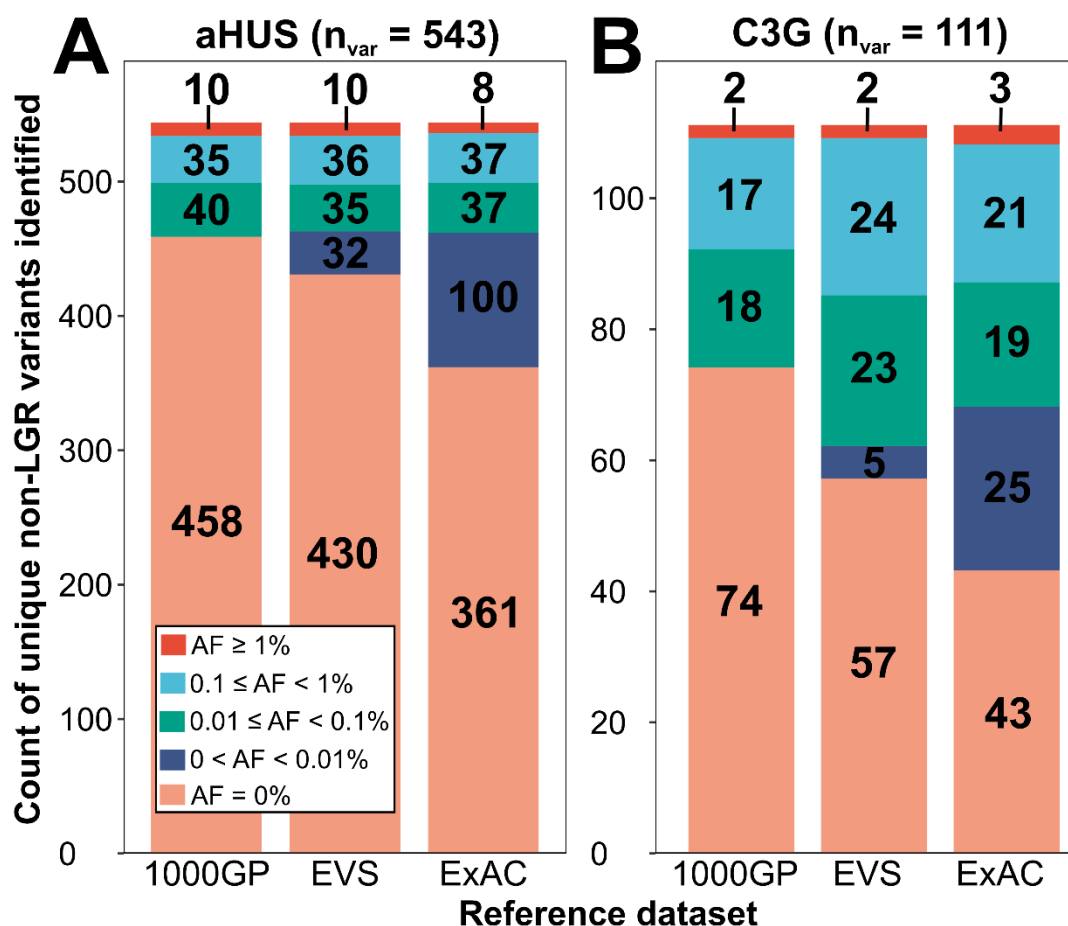


Figure 4.1 Stacked bar analyses showing the reference AF of the variants identified in the (A) aHUS and (B) C3G datasets.

The total numbers of unique variants (n_{var}) are shown within the bars. A unique variant is defined as, when a variant is seen in more than one patient, the variant is only counted once. Unique variants (excluding the 13 aHUS and 3 C3G copy number variants) are categorised by their AF in each of the three reference datasets: the 1000 Genomes Project (1000GP) with an allele number (AN; total number of alleles screened for each gene) of 3775-5008, the Exome Variant Server (EVS) with an AN of 8202-13005 and the Exome Aggregation Consortium (ExAC) with an AN of 14708-121412. The aHUS and C3G datasets each had ANs of 634-6256 and 208-886, respectively. The pink bars indicate a reference AF of 0%, dark blue bars indicate a reference AF of between 0% and 0.01% (non-inclusive), green bars indicate a reference AF of between 0.01% (inclusive) and 0.1%, light blue bars indicate a reference AF of between 0.1% (inclusive) and 1%, and red bars indicate a reference AF $\geq 1\%$.

Table 4.2 The 14 common genetic variants identified in at least one of the three reference datasets (1000GP, EVS and ExAC) at an AF of $\geq 1\%$. Bold text denotes statistical significance^{a,b}.

Gene	Genetic Variant	Protein Variant	Dataset Allele Frequency (%)				
			1000GP	EVS	ExAC	aHUS ^(a)	C3G ^(b)
<i>C3</i>	c.1407G>C	p.Glu469Asp	1.62^a	1.41^a	0.40	0.16	0
<i>CFB</i>	c.754G>A	p.Gly252Ser	1.02^a	2.82^a	2.22^a	0.43	0
<i>CFB</i>	c.1598A>G	p.Lys533Arg	1.92^{a,b}	0.44	1.05^a	0.20	0.26
<i>CFB</i>	c.1697A>C	p.Glu566Ala	1.02	0.73	1.12^a	0.53	0.13
<i>CFB</i>	c.1953T>G	p.Asp651Glu	1.04^a	0.94^a	0.22	0.04	0
<i>CFH</i>	c.1652T>C	p.Ile551Thr	1.91^a	1.68^a	0.50^a	0.16	0
<i>CFH</i>	c.2669G>T	p.Ser890Ile	6.23^a	6.57^a	1.99^a	0.34	0
<i>CFH</i>	c.2808G>T	p.Glu936Asp	20.33^a	13.80^a	19.55^a	0.02	0
<i>CFHR5</i>	c.136C>T	p.Pro46Ser	0.90	1.13	0.70	0.63	0
<i>CFHR5</i>	c.1067G>A	p.Arg356His	1.04	2.09^a	1.78^a	1.42	0.12
<i>CFI</i>	c.884-7T>C	-	2.56^a	0	2.31^a	0.02	0
<i>CFI</i>	c.1534+5G>T	-	0.30	1.13	0.90	0.33	0.12
<i>CD46</i>	c.1058C>T	p.Ala353Val	0.40	1.25^a	1.53^a	0.60	0.24
<i>PLG</i>	c.1567C>T	p.Arg523Trp	0.24	1.01	0.68	0.35	0

^a Determined to be significantly more common in ExAC than in aHUS using a two-tailed Chi-square test (χ^2) with Yates' correction and a Bonferroni-corrected significance level of 0.0036.

^b Determined to be significantly more common in ExAC than in C3G using a two-tailed Chi-square test (χ^2) with Yates' correction and a Bonferroni-corrected significance level of 0.0036.

threshold (for RVs) except for three (*CFH* p.Ser890Ile, *CFH* p.Glu936Asp and *CFI* c.884-7T>C). In the aHUS dataset, 14% (75), 19% (103), and 32% (174) of variants were found to have an AF between 0-1% (light blue/green/dark blue, Figure 4.1A) in the 1000GP, EVS and ExAC reference datasets, respectively. In the C3G dataset, these proportions were significantly higher, being 31% (35), 46% (52) and 58% (65) respectively (light blue/green/dark blue, Figure 4.1B). In confirmation of this outcome, all three analyses gave $p < 0.0001$ in the two-tailed Fisher's exact test. This outcome was still true when the 'benign' and 'likely benign' RVs were excluded from the aHUS and C3G datasets ($p < 0.0001$; Fisher's exact test) (see below). The aHUS and C3G RV datasets were compared with the literature (Rodriguez et al., 2014) in order to determine how many RVs were novel. Of the 542 aHUS and 110 C3G RVs, 56% (371) and 58% (82) respectively were not found in the literature and therefore novel (purple, Figure 4.2A). The RVs that were reported in the literature but not in the six Reference Centres were not part of my current analyses because I could not calculate their aHUS and C3G AF values (green, Figure 4.2A). However, they are included in the updated database from this study.

4.4.2 Rare variant frequencies in cases

The frequencies of the RVs in the aHUS and C3G datasets were surveyed. First, I investigated differences in the screened cases with RVs for aHUS and C3G. The aHUS and C3G datasets comprised 3128 aHUS and 443 C3G patients respectively (Figure 4.2B; Table 4.3). Of these totals, 1231 (39%) aHUS and 116 (26%) C3G patients harboured 542 and 110 unique RVs respectively, with an overlap of 42 unique variants between them. A significantly greater proportion of screened aHUS patients had at least one RV compared to C3G patients ($p < 0.0001$; two-tailed Fisher's exact test). There were 0.44 unique RVs per case in aHUS, compared to 0.95 in C3G. This suggested that almost every C3G case had a different, unique RV, whereas aHUS cases were more likely to share the same RV. The proportion of aHUS patients with RV (1231/3128=39.3%) (Figure 4.2B) was low compared to literature reports of ~60%. This difference most likely arose from the omission in the AF analyses of the 119 literature-sourced aHUS RVs in the web-database, for which no aHUS patient AF data from the six centres were available (green, Figure 4.2A).

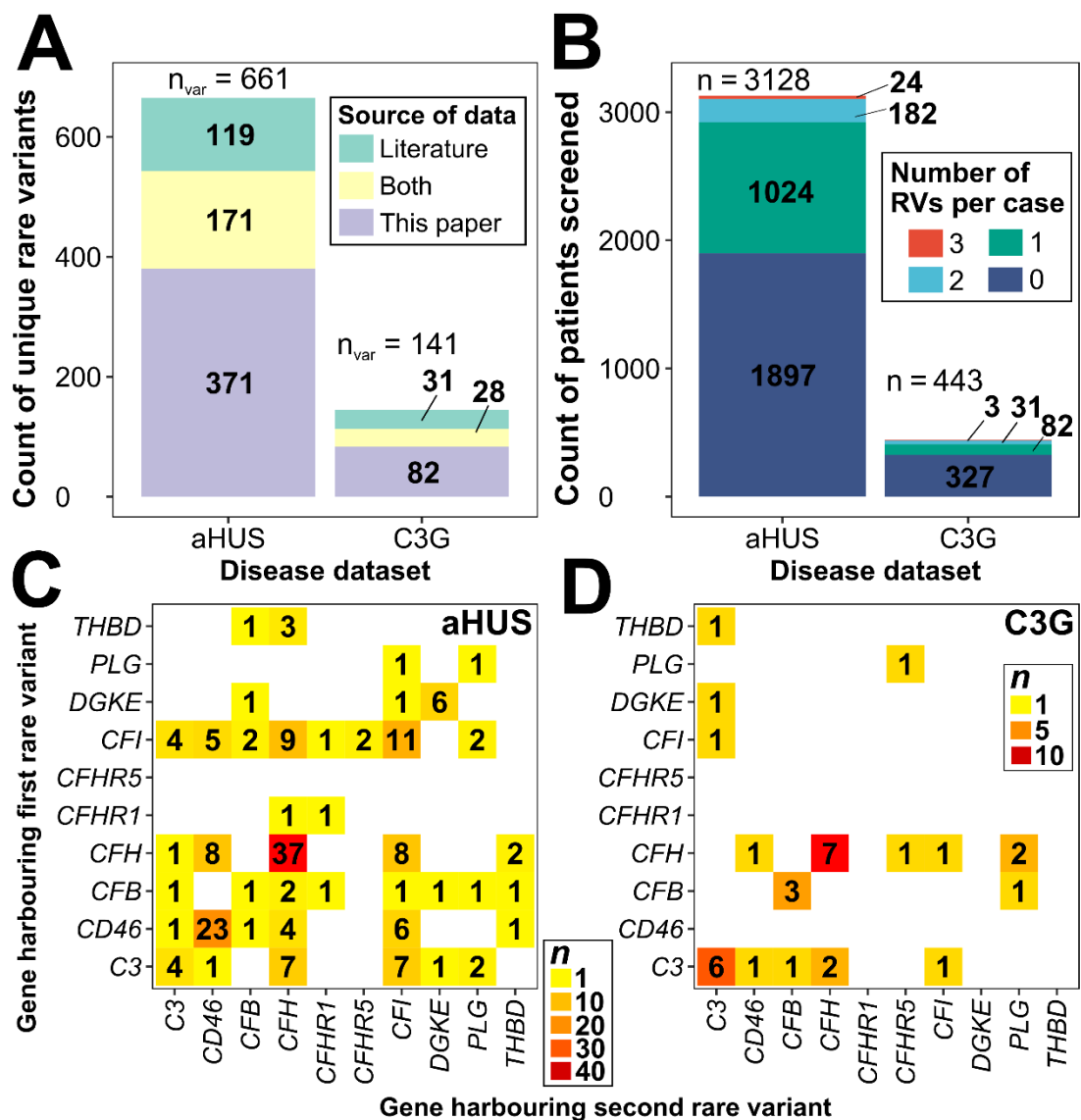


Figure 4.2 Summary of cases and variants in aHUS and C3G.

(A) The source of the RV data in the database for aHUS and C3G. “Both” (yellow) indicates RVs that were identified both in the laboratory-sourced datasets (purple) and published in the literature (green).

(B) The number of unique RVs (0 to 3) per patient case in the aHUS and C3G datasets, totalling 3127 patients. For aHUS, there is a further case with four RVs that is too small to be seen.

(C) A matrix showing the genetic profiles of the 182 aHUS cases with compound heterozygous (in single or in two different genes) or homozygous RVs from panel (B). Their frequencies n are graded in colour (inset).

(D) A matrix showing the genetic profiles of the 31 C3G cases with compound heterozygous (in single or in two different genes) or homozygous RVs from panel (B), graded in colour (inset).

Table 4.3 The total number of aHUS and C3G cases screened per gene

Gene	Disease	Total number of cases screened by the reference centres 1-6						
		1	2	3	4	5	6	All
<i>C3</i>	aHUS	410	480	286	252	409	618	2455
<i>CD46</i>	aHUS	524	480	286	461	578	613	2942
<i>CFB</i>	aHUS	395	480	286	328	350	618	2457
<i>CFH</i>	aHUS	662	480	286	483	578	639	3128
<i>CFHR1</i>	aHUS	7	250		442			699
<i>CFHR3</i>	aHUS		250		348			598
<i>CFHR5</i>	aHUS			286		31		317
<i>CFI</i>	aHUS	533	480	286	425	578	621	2923
<i>DGKE</i>	aHUS		76	286	191	150		703
<i>PLG</i>	aHUS			286				286
<i>THBD</i>	aHUS		480	286	348	214		1328
<i>C3</i>	C3G	115	160	104				379
<i>CD46</i>	C3G	142	160	104				406
<i>CFB</i>	C3G	115	160	104				379
<i>CFH</i>	C3G	179	160	104				443
<i>CFHR5</i>	C3G			104				104
<i>CFI</i>	C3G	144	160	104				408
<i>DGKE</i>	C3G		23	104				127
<i>PLG</i>	C3G			104				104
<i>THBD</i>	C3G		160	104				264

Centre 1: Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom

Centre 2: Clinical Research Center for Rare Diseases “Aldo e Cele Daccò”, IRCCS-Istituto di Ricerche Farmacologiche “Mario Negri”, Ranica Bergamo, Italy

Centre 3: Molecular Otolaryngology and Renal Research Laboratories, Carver College of Medicine, University of Iowa, Iowa City, Iowa, United States of America;

Centre 4: Department of Cellular and Molecular Medicine, Center for Biological Research and Center for Biomedical Network Research on Rare Diseases, Madrid, Spain;

Centre 5: Assistance Publique-Hopitaux de Paris, Hôpital Européen Georges Pompidou, Service d’Immunologie Biologique, Paris, France;

Centre 6: Department of Pediatric Nephrology, Radboud University Medical Center, Nijmegen, The Netherlands

4.4.3 Rare variant profiles of cases

Next, I compared the genetic RV profiles of aHUS and C3G patients. Among the 1231 aHUS and 116 C3G patients with RVs, 1024 (83%) and 82 (68%) respectively harboured a single RV in one of the 13 genes (Figure 4.2B). Two RVs were identified in 182 (15%) aHUS and 31 (29%) C3G cases (light blue bars; Figure 4.2B), three in 24 (2%) and 3 (2%) respectively (red bars; Figure 4.2B), and four in 1 (<1%) and 0 (0%) respectively. However, the number of RVs per aHUS case in the dataset from Reference Centre 6 was unknown except for three homozygotes. All other RV occurrences in this dataset were analysed as one heterozygous aHUS case. When the 322 aHUS cases from Reference Centre 6 were excluded, 702/909 (77%) harboured a single RV, 182/909 (20%) had two, 24/909 (3%) had three, and 1 (<1%) had four RVs. Excluding Reference Centre 6, no significant difference was seen between aHUS and C3G in the cases with two RVs ($p=0.113$; two-tailed Fisher's exact test). For aHUS, cases with two RVs in *CD46* (23 cases including 19 homozygotes; 11% of all aHUS cases with *CD46* RVs), were the most frequent, followed by *CFH* (37 cases, including 18 homozygotes; 9%) (Figure 4.2C; Table 4.4). For C3G, cases with two RVs in *CFH* (7 cases including 6 homozygotes; 30%) were the most frequent, followed by *C3* (6 cases including 2 homozygotes; 23%) (Figure 4.2D; Table 4.4). These analyses showed differences between aHUS and C3G, summarised in Figures 4.2C, 4.2D. No *CFHR3* RVs were seen in any aHUS or C3G cases that had more than one RV.

4.4.4 Gender analyses

Finally, I examined gender dependences in aHUS and C3G cases. Of the 1231 aHUS patients with at least one identified RV, 36% (440) were female, and 30% (363) were male (Table 4.5). In aHUS, assuming the remaining 34% of patients for which the gender was unknown showed a similar pattern, the bias towards females was significant ($p=0.007$). However, when *CFH* was removed from the analyses, the numbers of females (244) and males (225) was not significantly different ($p=0.38$) (Table 4.6). For the 116 C3G patients, no significant difference in the number of females (53) and males (63) was seen ($p=0.35$) (Table 4.6). In terms of gender differences, the aHUS dataset (542 variants) showed a bias towards females in *CFH*, which is likely to be a reflection of potential triggering factors. For example, in pregnancy-aHUS, most patients harbour *CFH* variants (Bruehl et al., 2017; Huerta et al., 2017). This trend was not observed for C3G.

Table 4.4 Homozygous RVs present in aHUS and C3G

Gene	cDNA change	Protein change	Classification	Disease	Cases
<i>C3</i>	c.3322_3333del	p.1108_1111del	Likely benign	aHUS	1
<i>CD46</i>	c.97G>C	p.Asp33His	Uncertain significance	aHUS	2
<i>CD46</i>	c.100G>A	p.Ala34Thr	Uncertain significance	aHUS	1
<i>CD46</i>	c.286+2T>G	-	Pathogenic	aHUS	4
<i>CD46</i>	c.286+1G>C	-	Pathogenic	aHUS	1
<i>CD46</i>	c.496T>C	p.Cys157Arg	Uncertain significance	aHUS	1
<i>CD46</i>	c.535G>C	p.Glu179Gln	Pathogenic	aHUS	1
<i>CD46</i>	c.565T>G	p.Tyr189Asp	Likely pathogenic	aHUS	1
<i>CD46</i>	c.718T>C	p.Ser240Pro	Pathogenic	aHUS	2
<i>CD46</i>	c.736T>A	p.Phe246Ile	Uncertain significance	aHUS	1
<i>CD46</i>	c.800_820del	-	Uncertain significance	aHUS	1
<i>CD46</i>	c.810T>G	p.Cys270Trp	Uncertain significance	aHUS	1
<i>CD46</i>	c.811_816delGACAGT	p.Asp271_Ser272del	Pathogenic	aHUS	1
<i>CD46</i>	c.881C>T	p.Pro294Leu	Uncertain significance	aHUS	1
<i>CD46</i>	c.1027+2T>C	-	Pathogenic	aHUS	1
<i>CFH</i>	c.79_82delAGAA	-	Likely pathogenic	aHUS	1
<i>CFH</i>	c.157C>T	p.Arg53Cys	Pathogenic	aHUS	2
<i>CFH</i>	c.158G>A	p.Arg53His	Likely pathogenic	aHUS	1
<i>CFH</i>	c.269T>A	p.Tyr899*	Likely pathogenic	aHUS	2
<i>CFH</i>	c.2880delT	p.Phe960fs	Likely pathogenic	aHUS	1
<i>CFH</i>	c.2918G>A	p.Cys973Tyr	Uncertain significance	aHUS	1
<i>CFH</i>	c.3048C>A	p.Tyr1016*	Likely pathogenic	aHUS	2
<i>CFH</i>	c.3628C>T	p.Arg1210Cys	Pathogenic	aHUS	1
<i>CFH</i>	[c.3674A>T; 3675_3699del]	p.Tyr1225Tyrfs*38	Likely pathogenic	aHUS	2
<i>CFH</i>	c.3693_3696 delATAG	p.*1232Ilefs*38	Likely pathogenic	aHUS	5
<i>CFI</i>	c.341T>C	p.Val114Ala	Uncertain significance	aHUS	1
<i>CFI</i>	c.1357T>C	p.Cys453Arg	Uncertain significance	aHUS	1
<i>CFI</i>	c.1456T>C	p.Trp486Arg	Uncertain significance	aHUS	1
<i>CFI</i>	c.1642G>C	p.Glu548Gln	Uncertain significance	aHUS	2
<i>DGKE</i>	c.889-1G>A	p.IVS5-1	Likely pathogenic	aHUS	1
<i>DGKE</i>	c.966G>A	p.Trp322*	Likely pathogenic	aHUS	4
<i>DGKE</i>	c.1000C>T	p.Gln334*	Likely pathogenic	aHUS	1
<i>DGKE</i>	c.1608_1609del	p.His536Glnfs*16	Likely pathogenic	aHUS	1
<i>PLG</i>	c.1481C>T	p.Ala494Val	Likely benign	aHUS	1
<i>C3</i>	c.168_169delTG	p.Thr56Thrfs*16	Uncertain significance	C3G	1
<i>C3</i>	c.1682G>A	p.Gly561Asp	Uncertain significance	C3G	1
<i>CFB</i>	c.2035C>T	p.Arg679Trp	Uncertain significance	C3G	1
<i>CFH</i>	c.232A>G	p.Arg78Gly	Pathogenic	C3G	1
<i>CFH</i>	c.262C>A	p.Pro88Thr	Uncertain significance	C3G	1
<i>CFH</i>	c.694C>T	p.Arg232*	Likely pathogenic	C3G	1
<i>CFH</i>	c.3286T>A	p.Trp1096Arg	Uncertain significance	C3G	3

* Denotes a nonsense variant

Table 4.5 Demographics of the 1231 aHUS and 116 C3G cases showing an identified RV

Disease	Gender	Number of cases aged 18+ years ^a	Number of cases aged <18 years ^b	YOB ^c unknown	Total
aHUS	Female	328 (27%)	84 (7%)	28 (2%)	440 (36%)
aHUS	Male	254 (20%)	84 (7%)	25 (2%)	363 (29%)
aHUS	Unknown	2 (<1%)	1 (<1%)	425 (34%)	428 (35%)
aHUS	ALL	584 (48%)	169 (14%)	478 (38%)	1231
C3G	Female	34 (29%)	12 (10%)	7 (6%)	53 (46%)
C3G	Male	40 (35%)	16 (14%)	7 (6%)	63 (54%)
C3G	Unknown	0	0	0	0
C3G	ALL	74 (64%)	28 (24%)	14 (12%)	116

^a Born in or before 1997

^b Born after 1997

^c Year of birth

Table 4.6 Gender of the 1231 aHUS and 116 C3G cases with an identified RV

Disease	Gene	Female	Male	Unknown	Proportion of females (known) (%)	<i>P</i> ^a
aHUS	<i>C3</i>	76	61	99	55	0.20
aHUS	<i>CD46</i>	77	91	67	46	0.28
aHUS	<i>CFB</i>	13	11	11	54	0.68
aHUS	<i>CFH</i>	196	138	196	59	0.002
aHUS	<i>CFHR1</i>	8	13	1	38	0.28
aHUS	<i>CFHR3</i>	1	1	0	50	1.00
aHUS	<i>CFHR5</i>	4	3	0	57	0.71
aHUS	<i>CFI</i>	80	60	54	57	0.09
aHUS	<i>CFP</i>	1	1	0	50	1.00
aHUS	<i>DGKE</i>	16	9	0	64	0.16
aHUS	<i>PLG</i>	7	5	0	58	0.56
aHUS	<i>THBD</i>	16	8	10	67	0.10
aHUS	ALL ^b	440	363	428	55	0.007
C3G	<i>C3</i>	23	20	0	53	0.65
C3G	<i>CD46</i>	1	1	0	50	1.00
C3G	<i>CFB</i>	4	4	0	50	1.00
C3G	<i>CFH</i>	19	24	0	44	0.45
C3G	<i>CFHR1</i>	0	1	0	0	0.32
C3G	<i>CFHR3</i>	0	1	0	0	0.32
C3G	<i>CFHR5</i>	0	4	0	0	0.05
C3G	<i>CFI</i>	5	6	0	45	0.76
C3G	<i>DGKE</i>	1	2	0	33	0.56
C3G	<i>PLG</i>	2	5	0	29	0.26
C3G	<i>THBD</i>	2	7	0	22	0.10
C3G	ALL ^b	53	63	0	46	0.35

^a P value from a two-tailed Chi-square test (χ^2) with Yates' correction using a Bonferroni-corrected significance level of 0.0042 for aHUS and 0.0045 for C3G. The 'ALL' genes category was subject to a significance level of 0.05 (no Bonferroni correction). Bold text denotes a P level less than the Bonferroni-corrected significance level.

^b 'ALL' is not equal to the gene sum, because some patients possessed RVs in more than one unique gene.

4.4.5 Rare variant pathogenicity classification

RVs were categorised for their pathogenicity using published experimental evidence, reference AFs and *in silico* predictive analyses, in accordance with genetic variant guidelines (Goodship et al., 2017) (Figure 4.3A). The average healthy genome also contains benign RVs that occur at similar AFs to disease-related RVs, therefore AF analyses alone cannot be used as evidence for pathogenicity (Kobayashi et al., 2017). In terms of pathogenicity, only gain-of-function RVs in the activators *C3* and *CFB* are expected to predispose for aHUS or C3G disease; here gain-of-function results from the loss of ability to interact with an inhibitory complement regulator. The predictive tools PolyPhen-2 and SIFT were unable to predict gain-of-function phenotypes, thus many *C3* and *CFB* RVs without experimental data could only be classified as ‘uncertain significance’ (green bars; Figure 4.3A). For example, the *C3* RV p.Arg161Trp was predicted to be damaging and deleterious, while functional studies showed a *C3* gain-of-function (Roumenina et al., 2012). The number of ‘pathogenic’ or ‘likely pathogenic’ RVs in *C3* and *CFB* were therefore lower than for the other genes. In aHUS, the majority of RVs in *CFH*, *CFI*, *CD46* and *DGKE* were either ‘pathogenic’ or ‘likely pathogenic’ (red and light blue bars; Figure 4.3A), this being attributed to their loss-of-function phenotypes. In C3G, the majority of RVs in all genes were of ‘uncertain significance’ (green bars; Figure 4.3A) except for *CFH*. In terms of the non-pathogenic RVs in the aHUS and C3G datasets, a survey of all 542 RVs in aHUS and 110 in C3G showed that 43 (8%) and 27 (25%) RVs respectively were categorised as ‘benign’ and ‘likely benign’, and these were filtered out (Appendix I). This suggested that the RVs in the C3G dataset may be less pathogenic compared to aHUS overall; this observation reinforced the importance of using classification guidelines prior to the protein domain hotspot analyses below. I also note that ‘likely pathogenic’ RVs still require further functional studies to confirm or disprove the predicted effects and disease relevance of the variants.

4.4.6 Rare variant abundance in genes

In order to identify differences in the frequency of unique RVs per gene between aHUS and C3G, and therefore the molecular pathogenesis of both diseases, the abundances of RVs in each gene were analysed. I stress that only those RVs that were not classified as ‘benign’ or ‘likely benign’ were analysed (red, light blue, green and grey bars; Figure 4.3A). The *CFHR1-CFHR4* genes were subject to multiple

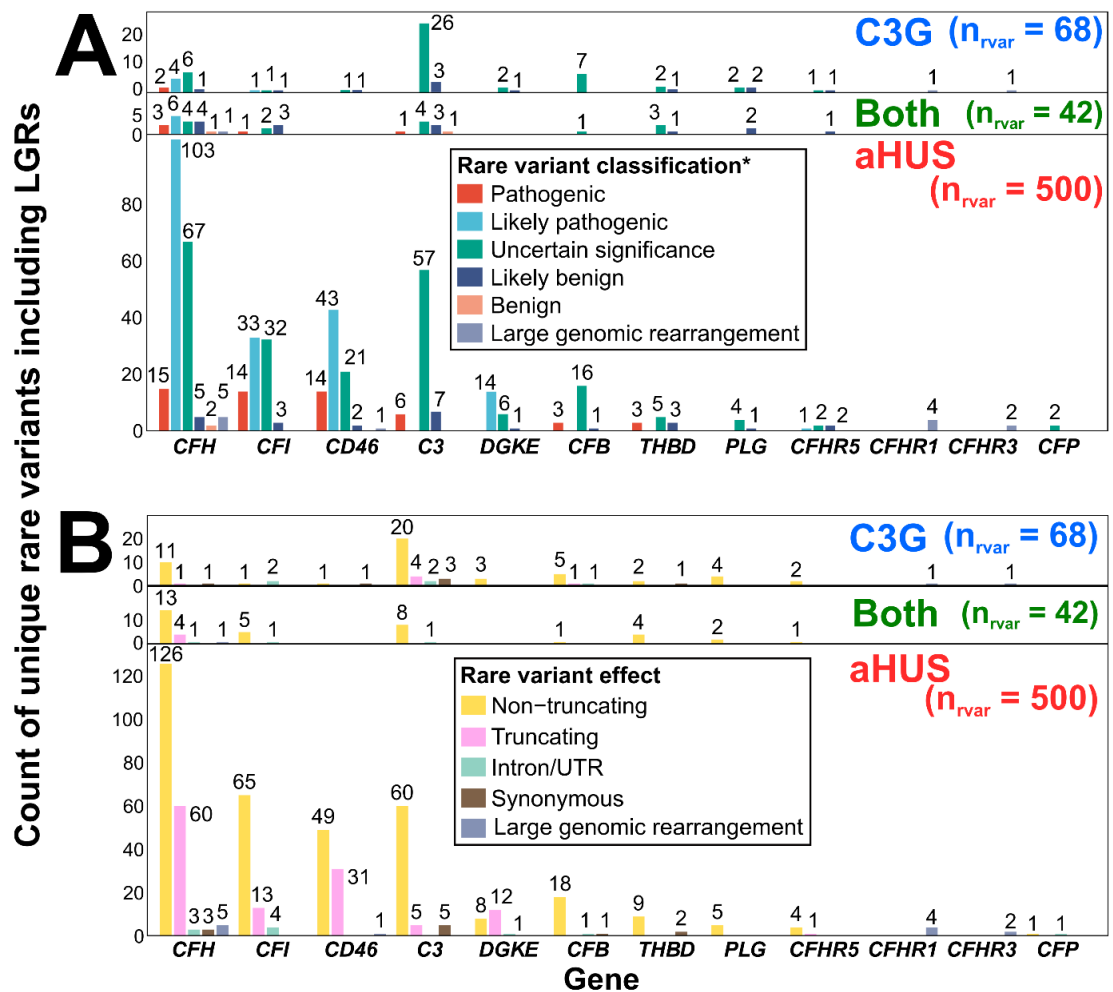


Figure 4.3 RV effects and classifications in aHUS and C3G. The colour coding of each RV effect and its classification is shown in the insets.

(A) In terms of pathogenicity, the total number of unique RVs for each gene and their classification, based on the pathology guidelines (Methods), are shown for the aHUS and C3G datasets, and for both datasets, in three panels.

(B) In terms of functional annotation (such as that used in ExAC), the total number of unique RVs for each gene and their effect on each protein are shown for the aHUS and C3G datasets, and for both datasets in three panels.

ligation-dependent probe assessment only, and were not sequenced, thus only large genomic rearrangements were identified. In aHUS, *CFH* showed the most RVs (204 variants; 41%) (Figure 4.3A), followed by *CFI* (82; 17%), *CD46* (79; 16%), and *C3* (68; 14%). This outcome confirmed the 2014 analyses for *CFH*, *CFI*, *CD46* and *C3* (Rodriguez et al., 2014). Lesser abundances for aHUS were seen for *CFB* (20; 4%), *DGKE* (20; 4%); *THBD* (11; 2%), *PLG* (4; 1%), *CFHR5* (3; 1%), *CFHR1* (4; 1%), *CFHR3* (2; <1%) and *CFP* (2; <1%). Only one unique variant in *CFHR4*, a *CFHR1/CFHR4* deletion, was seen in three heterozygous and one homozygous aHUS cases. This was classified under the *CFHR1* gene in the database. In C3G, *C3* (31 variants; 37%) and *CFH* (25 variants; 28%) showed the most RVs (red, light blue, green and grey bars; Figure 4.3A). Lesser abundances for C3G were seen for *CFI* (5; 6%), *CD46* (1; 1%), *CFB* (8; 10%), *DGKE* (2; 2%); *THBD* (2; 2%), *PLG* (2; 2%), *CFHR5* (1; 1%), *CFHR1* (1; 1%), *CFHR3* (1; 1%) and *CFP* (0; 0%). It was concluded that RVs in *CFI* and *CD46* were substantially reduced in C3G compared to aHUS ($p=0.0121$ and $p<0.0001$, respectively; two-tailed Fisher's exact test). When analysed in terms of genetic effect for both aHUS and C3G, most RVs were non-truncating (yellow; Figure 4.3B), except for the aHUS variants in *DGKE* that were mostly truncating, and the large genomic rearrangements in *CFHR1* and *CFHR3*. The most frequent RV in the aHUS dataset was *C3* p.Arg161Trp, seen at an AF of 1.16% in 52 aHUS cases. However in C3G, none of the RVs were notably more frequent than others. Data on RVs found in the complement inhibitor vitronectin and clusterin genes in the aHUS and C3G datasets were not available for analysis at the time of writing (Stahl et al., 2009; van den Heuvel et al., 2018).

4.4.7 Gene-based rare variant burden

In order to confirm that the amount of rare variation seen in the genes of aHUS and C3G patients was greater than in the genes of individuals without these diseases, I determined the burden of protein-altering rare variation (ExAC MAF<0.01%) per gene for each dataset (Methods). These were compared to the ExAC and EVS reference datasets (Figure 4.4; Appendix II and III). Because ExAC and EVS did not contain data on large genomic rearrangements, *CFHR1-4* were not analysed. No aHUS or C3G AF data were available for *CFP*. This left nine out of 13 genes (*CFH*, *CFI*, *CD46*, *C3*, *DGKE*, *CFB*, *CFHR5*, *PLG*, *THBD*) for analysis. For the aHUS dataset, a significantly greater burden of rare variation was revealed in patients than in the ExAC and EVS datasets for five genes, namely *CFH*, *CFI*, *CD46*, *C3* and *DGKE* (Chi-square test (χ^2) with Yates'

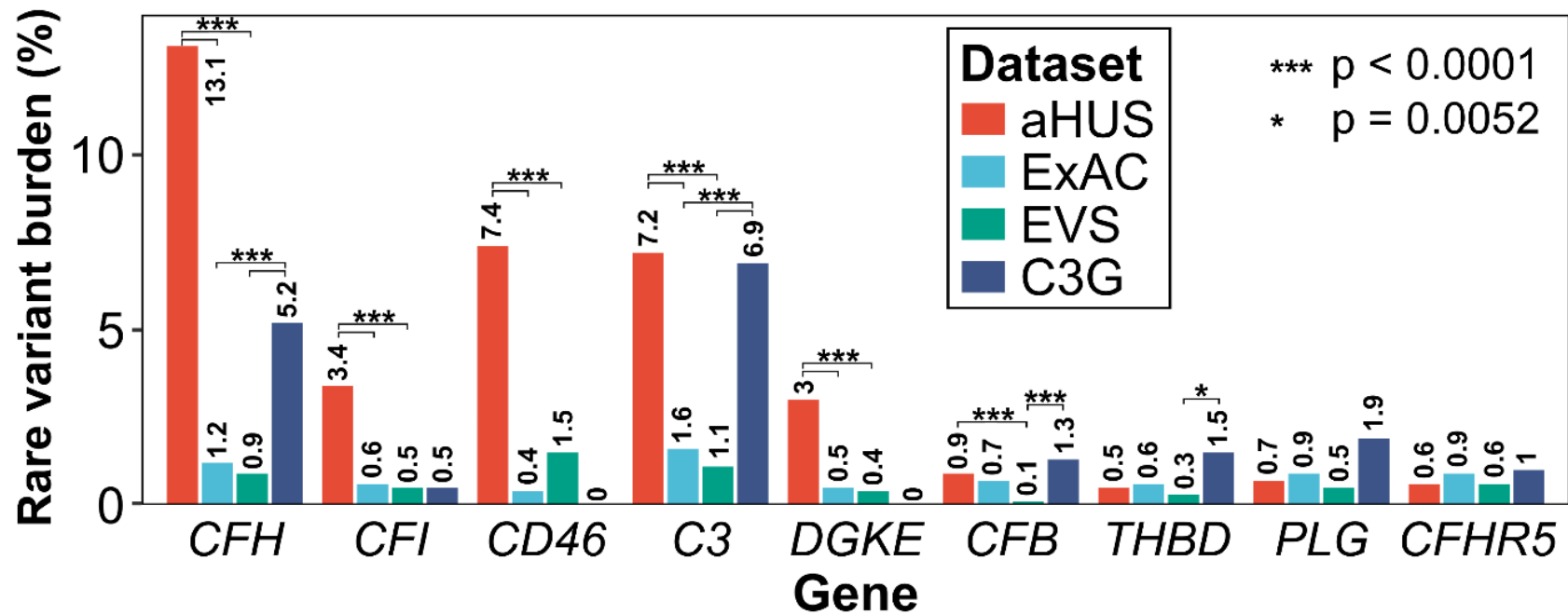


Figure 4.4 The RV burden (%) per gene for the nine relevant genes in the four aHUS (Allele number (AN): 634-6256), ExAC (AN: 74194-121246), EVS (AN: 8202-13005) and C3G (AN: 208-886) datasets. These were based on an ExAC MAF cut-off of 0.01%. *** denotes $p < 0.0001$. * denotes $p = 0.0052$.

correction using a Bonferroni-corrected significance level of 0.0056; the five genes each gave $p < 0.0001$), and also for *CFB* when compared with the EVS dataset only (also with $p < 0.0001$) (Figure 4.4; Appendix II and III). No association with aHUS was observed with RVs in *THBD*, *PLG* and *CFHR5*. The tests for these three genes showed a power $> 80\%$ thus their false negative rates were expected to be very low. In the C3G dataset, *C3* ($p < 0.0001$; $p < 0.0001$) and *CFH* ($p < 0.0001$; $p < 0.0001$) showed a significantly greater burden of protein-altering rare variation in patients than in the ExAC and EVS datasets (Chi-square test (χ^2) with Yates' correction using a Bonferroni-corrected significance level of 0.0056) (Figure 4.4; Appendix II and III). C3G was also associated with RVs in *CFB* ($p < 0.0001$) and *THBD* ($p = 0.0052$) when only EVS was used as the reference dataset (green bar, Figure 4.4; Appendix III), despite both tests showing a lack of power (*CFB*: 41% and *THBD*: 38%). The lack of association of *DGKE* and *PLG* with C3G may also relate to lack of power (*DGKE*: 55% ExAC and 68% EVS, and *PLG*: 32% ExAC and 10% EVS) shown in the tests for these genes.

4.4.8 Distribution of aHUS and C3G RVs in FH

The aHUS and C3G location and AF of the missense RVs for each gene resulted in the identification of mutational hotspots in each protein structure. For FH, as seen in the 2014 study (Rodriguez et al., 2014), most aHUS missense RVs (78) occurred in the C-terminal ten domains, compared to the N-terminal ten domains (47; $p = 0.0042$) (Figure 4.5A; Appendix IV). In aHUS, the total frequency of *CFH* alleles with a missense RV in the C-terminal ten domains (3.2%) was significantly greater than for the N-terminal ten domains (1.2%; Chi-square test with Yates' correction using a significance level of 0.05; $p < 0.0001$). In SCR-20, the total missense RV aHUS AF of 2.03% (Appendix IV) was the highest for all 20 domains. These non-random distributions supported the functional association of SCR-20 with cell surface dysregulation in aHUS, although laboratory experimentation and functional characterization will be required to validate this result. This was still the case when normalised for the size of the FH domain (62/1231 residues; 5.5%), giving 37.2% (Figure 4.5A; Appendix IV). Most of the six FH domains (SCR-2, SCR-5, SCR-8, SCR-12, and SCR-13) with the least number of aHUS RVs did not correspond to known FH binding sites. For C3G, in contrast to aHUS, the C3G missense RVs in FH were clustered at the N-terminal C3b binding site, with comparatively few at the non-surface associated C-terminal domains such as SCR-15 (Figure 4.5A). No C3G clusters were seen in the heparin binding regions of FH. The only three unique C3G

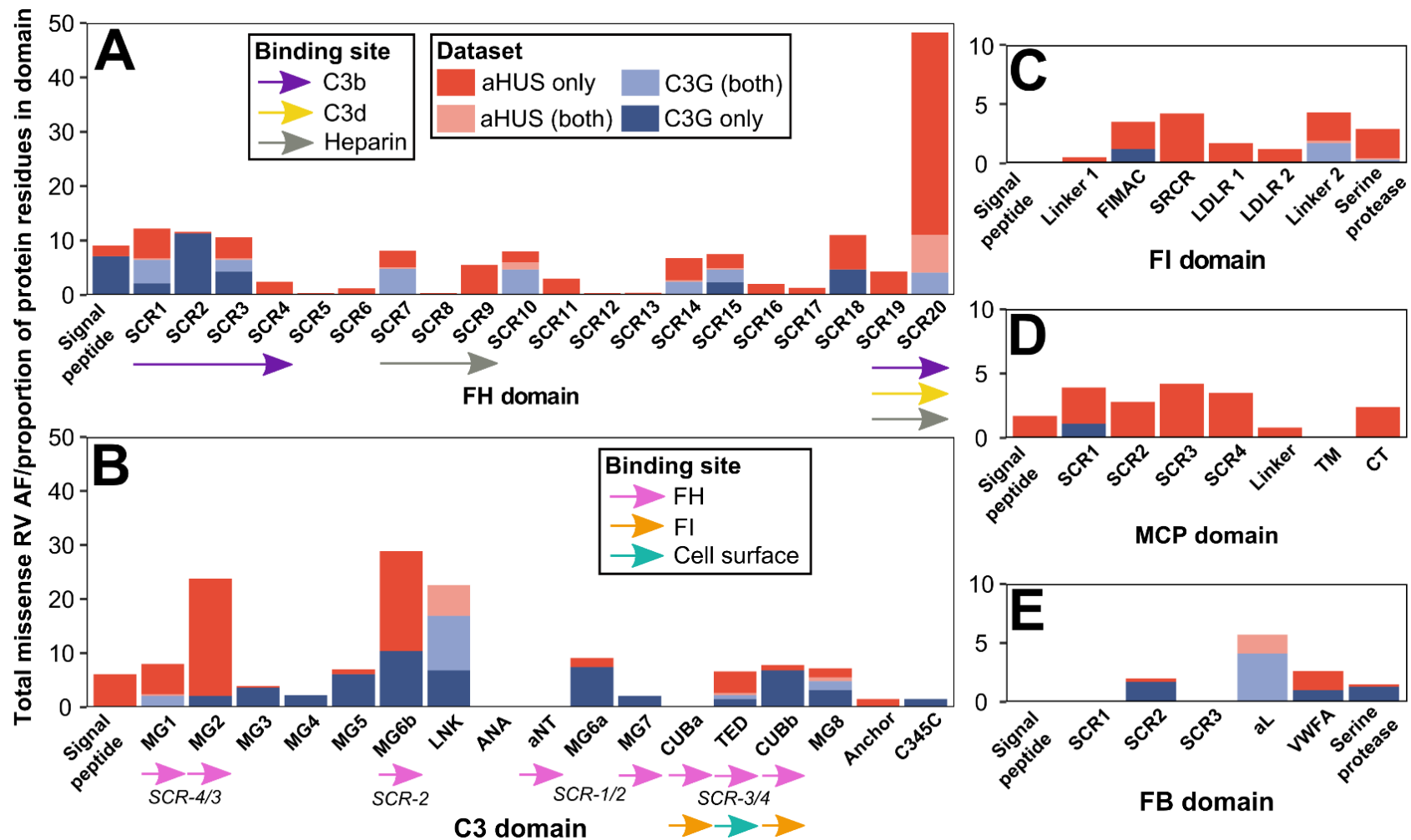


Figure 4.5 The distribution and disease allele frequencies (AFs) of non-benign missense RVs in the domains of FH, C3, FI, MCP, and FB in the aHUS and C3G datasets. Figure legend overleaf.

Figure 4.5 (continued) The distribution and disease allele frequencies (AFs) of non-benign missense RVs in the domains of FH, C3, FI, MCP, and FB in the aHUS and C3G datasets. The largest bars correspond to missense RV hotspots (e.g. FH SCR-20 for aHUS; SCR-18 for C3G). Each domain missense RV AF is normalised for its size by dividing it by the proportion of residues in the protein domain. In each of (A-E), red represents the total AF missense RVs identified in the aHUS dataset, and likewise dark blue for C3G. For those missense RVs identified in both the aHUS and C3G datasets, pink represents the AF for aHUS and light blue for C3G. On the x-axes, the domain names are shown.

(A) The total AF of missense RVs in each of the 20 SCR domains in FH. Beneath the x-axis, the functional binding sites associated with each SCR domain are shown by coloured arrows (identified in the inset).

(B) The total AF of missense RVs in each C3 domain. Beneath the x-axis, the functional binding sites associated with each C3 domain are shown by arrows to correspond to the SCR domains in FH (pink) or other sites in FI or on the cell surface. The C3d binding site on SCR-19/20 corresponds to the TED domain, however this is not shown.

(C) The total AF of missense RVs in each FI domain.

(D) The total AF of missense RVs in each MCP domain.

(E) The total AF of missense RVs in each FB domain.

missense RVs found in SCR-7 (p.Cys431Tyr; C3G AF: 0.2%) and SCR-20 (p.Arg1210Cys, 0.1%; p.Cys1218Arg, 0.1%) in the heparin binding regions were also seen in aHUS at similar AFs of 0.01%, 0.4%, and 0.01% respectively, suggesting an overlap in phenotypes. The C3G missense RVs in FH were also not found in eleven SCR domains (SCR-4/6, SCR-8/9, SCR-11/13, SCR-16/17 and SCR-19). There was a C3G missense RV cluster at the N-terminal SCR-1/3 C3b binding region (Thomas et al., 2014) (Figure 4.5A; Appendix IV); I infer that SCR-1/3 may be a mutational hotspot for C3G. In summary, the distribution of FH hotspots between aHUS and C3G showed clear differences between the two diseases.

4.4.9 Distribution of aHUS and C3G rare variants in C3

For C3, the 67 aHUS C3 missense RVs occurred in 12 of the 16 C3 domains (Figure 4.5B; Appendix IV). Macroglobulin (MG)-2 showed the highest missense RV AF for C3 (1.3%), normalised by the proportion of domain residues (21.7%), followed by MG-6b (0.5%; 18.5%) (Figure 4.5B). Both the MG-2 and MG-6b domains were thus inferred to be aHUS hotspots. The 29 C3G missense RVs occurred in 13 of the 16 C3 domains (Appendix IV). In contrast to aHUS, these were spread more evenly throughout C3 and no hotspots were inferred (Figure 4.5B). No aHUS or C3G variants involved the key C3 thioester residues (Glu991, Cys998, His1104 and Glu1106), or the anaphylatoxin (ANA), aNT, and beta-sheet (CUB)-a domains. Again the distribution of C3 missense RVs between aHUS and C3G showed clear differences between the two diseases.

4.4.10 Distribution of aHUS and C3G rare variants in FI, CD46, FB and others

For FI, the 65 missense RVs in aHUS were distributed across all five domains of FI (Figure 4.5C). For MCP, the 45 aHUS rare missense variants were distributed across the four MCP domains (Figure 4.5D). For FB, there were no aHUS or C3G missense RVs in SCR-1 or SCR-3 (Figure 4.5E). In FB, most aHUS missense RVs occurred in the VWFA domain, whereas the C3G missense RVs were spread across the SCR-2, VWFA and serine protease domains. No mutational hotspots were evident for either aHUS or C3G in FI, MCP or FB.

4.4.11 Minor allele frequency analyses

The AFs of each of the protein-altering RVs per gene in the aHUS and C3G datasets were compared with the corresponding AF in ExAC. The two *CFP* RVs in aHUS were not analysed because there were no available patient AF data. *CFHR1*, *CFHR3* and *CFHR4* were not analysed because large genomic rearrangements were not included in ExAC at the time of writing. There were no variants that were rare in ExAC (MAF<0.01%) yet were significantly more common in ExAC, than in aHUS or C3G. The AFs and their significance are shown on the database for users.

4.5 Discussion

4.5.1 Summary of rare variants in aHUS and C3G

Compared to a 2014 study ([Rodriguez et al., 2014](#)) in which 324 aHUS- and C3G-associated genetic variants in *CFI*, *CFH*, *C3* and *CD46* were used to identify variant hotspots in these four proteins (www.fh-hus.org), my more detailed study now reports 610 RVs in 13 genes from 3128 aHUS and 443 C3G patients, together with their associated AFs. The expansion to 13 genes reflects new candidate genes for potential association with these diseases, and clarifies the extent to which they are indeed associated. To my knowledge, this is the largest study of aHUS and C3G patients to date. The much increased totals of RVs and cases now make possible for the first time AF analyses and RV burden analyses of the aHUS and C3G datasets, including comparisons to three reference genome datasets. In turn, the AF analyses provided more detailed statistical analyses of the RV distributions in both diseases. The analyses of distinct protein domain hotspots for aHUS and C3G clarify molecular differences that rationalise the occurrence of their two different phenotypes, the involvement of the RVs that are present, and the molecular mechanisms involved in both diseases. In particular, this study has reduced the earlier knowledge gap in the genetics and genotype-phenotype correlations of C3G to bring these closer to that of aHUS ([Goodship et al., 2017](#)).

4.5.2 Differences between aHUS and C3G using allele frequency analyses

The AF analyses revealed three new insights into the individual RVs associated with aHUS and C3G (and not the full genes). My analyses raised the question of what AF

cut-off to employ. Firstly, my AF analyses verified the rarity of 97% of the aHUS and C3G variants when compared to the ExAC reference (Figure 4.1), especially given the ability of ExAC to resolve ultra-RV AFs as low as 1×10^{-5} (0.001%) (Walsh et al., 2016). Such disease-predisposing alleles in aHUS and C3G are by definition deleterious. In theory, evolutionary pressures will maintain these alleles at very rare frequencies in the general population through negative selection (Gibson, 2011; Rallapalli et al., 2014; Zuk et al., 2014). Other studies report a histogram of variants in which pathogenicity increases with rarity and low AF (Kobayashi et al., 2017). These results therefore justify my focus on RVs with AF < 1% in the reference datasets. A more stringent RV AF cut-off of 0.01% is applicable to rare diseases of Mendelian inheritance (Kobayashi et al., 2017). This 0.01% cut-off was used for the RV burden calculations in order to restrict them to RVs with a higher confidence of pathogenicity (Figure 4.4; Appendix II and III). However, RVs observed in both aHUS and C3G have reduced penetrance, where disease-free individuals also harbour these pathogenic RVs, and the onset of aHUS or C3G disease depends upon a trigger and other factors. To enable my hotspot analyses, a less stringent reference AF cut-off of < 0.1% was thus used, but accompanied alongside experimental data and prediction tools, as specified in genetic variant classification guidelines¹. Some variants show AFs of > 1% (Table 4.2). While these may be risk factors for aHUS (e.g. *CFH* p.Val62Ile) (Noris & Remuzzi, 2017), their analysis was beyond the scope of this study. All the variants are available to view in the Database of Complement Gene Variants by adjusting the value of the ExAC AF filter.

The second insight involved AF differences between aHUS and C3G that relate to their different phenotypes. Significantly greater proportions of C3G RVs were identified in the reference datasets, with AFs between 0-1% rather than 0%, unlike aHUS (Figure 4.1). This meant that the C3G RVs occurred more frequently in individuals without C3G (the reference datasets) than the aHUS RVs. This result may explain the mostly chronic presentation of C3G that accumulates over time, in distinction to the mostly acute presentation of aHUS.

Thirdly, the AF analyses of the disease datasets (i.e. not the reference datasets) revealed differences between the most common RVs in the aHUS and C3G datasets that were likely to reflect their different phenotypes. In the aHUS dataset, while *CFH* had the most RVs, the most frequent RV was p.Arg161Trp in *C3*, corresponding to an AF of 1.16% of 3128 aHUS cases, i.e. 52 aHUS cases. *C3* p.Arg161Trp has a surface exposed

position in the MG2 domain and forms a hyperactive C3 convertase with an increased affinity for factor B, thus leading to over-activation of the AP. C3 p.Arg161Trp was not seen in the reference datasets, and classification guidelines confirmed its pathogenicity (Goodship et al., 2017). In the C3G dataset, C3 had the most RVs. Further analyses (below) reveal distinct domain hotspots for aHUS and C3G.

4.5.3 Rare variant burden testing

The RV burden is the proportion of screened cases with an identified protein-altering RV for which the ExAC AF was <0.01%. As opposed to the AF analyses that look at each variant one-by-one, the RV burden now provides insight into all the RVs associated with each gene. In general, RV burden tests assume that all tested RVs influence the phenotype in the same direction (Auer & Lettre, 2015). I therefore separated RVs into ‘truncating’ (loss-of-function) and ‘non-truncating’ (either loss- or gain-of-function, or neutral) in order to aid interpretations. RV burden tests showed clear differences between aHUS and C3G when compared to reference datasets, and this clarified the molecular mechanisms of the two diseases. Previous knowledge of experimental functional characterization of some of the RVs collectively support that aHUS is more related to surface AP dysregulation, while C3G is more related to fluid phase AP dysregulation. Here I now extend these earlier functional results:

(i) For aHUS, my RV burden analyses confirmed the association of rare variation in the six genes *CFH*, *CFI*, *CD46*, *DGKE* (Azzi et al., 1992; Lemaire et al., 2013), *C3* and *CFB*. For the five genes of the AP, *CFH*, *CFI*, *C3* and *CFB* are involved in both cell surface and fluid phase regulation, but *CD46* is only involved in cell surface regulation. Therefore, my RV burden analyses suggest that aHUS involves defects that result in both cell surface and fluid phase dysregulation. Different domains of C3 and FH are involved in cell surface dysregulation compared to fluid phase dysregulation, and this is explored in terms of the aHUS and C3G RV distributions in the next section (Section 4.5.4). The protein encoded by the sixth gene, *DGKE*, is found in endothelium, platelets and podocytes, and normally inactivates the signalling of arachidonic acid-containing diacylglycerols which activate protein kinase C (PKC) and promote thrombosis. Loss of *DGKE* function may thus result in a prothrombotic state and lead to microangiopathic haemolytic anaemia seen in aHUS outside the complement system (Lemaire et al., 2013). For three more genes *CFHR5*, *PLG* and *THBD*, no association of RVs in these with aHUS

was observed. Each of their tests showed a high power of >80%. This is unexpected from their known function, where FHR5 is likely to compete with FH for regulation (Csincsi et al., 2015; Goicoechea de Jorge et al., 2013), while PLG and THBD are inhibitors of thrombosis, and THBD also regulates complement (Conway, 2012; Delvaeye et al., 2009; Heurich et al., 2016; Wang et al., 2012).

(ii) For C3G in contrast, the RV burden analyses showed that the four genes *C3*, *CFH*, *CFB* and *THBD* were associated with C3G, while the two genes *CD46* and *CFI* were not associated with C3G. This outcome suggested that C3G is not caused by defects in cell surface regulation by *CD46* or defects in cell surface or fluid phase regulation by FI. My results also suggested that non-large genomic rearrangement RVs in *CFHR5* and *PLG* are not causative for C3G. I did not analyse large genomic rearrangements in *CFHR5* such as those identified in CFHR5 glomerulopathy. Despite pathogenic RVs in known cardiac genes not being overrepresented in ExAC (Song et al., 2016), *THBD* is involved in cardiac disease cases. Furthermore, the prevalence of RVs may differ across different centres, especially for genes such as *CFHR5* and *PLG* in aHUS, and *CFHR5*, *PLG* and *DGKE* in C3G that are less mutated and/or were sequenced by one or two centres only. This outcome is potentially affected by sampling bias, thus being difficult to interpret, and the results for these genes should be considered as preliminary only. In addition, the lack of association of *DGKE* and *PLG* with C3G may also be related to a lack of power (10% - 68%) shown by their tests.

4.5.4 Hotspots for missense rare variants

The 2014 identification of hotspots in four complement proteins (Rodriguez et al., 2014) can now be expanded to examine clusters of missense RVs in each of aHUS and C3G. Based on the RVs with 0.1% reference AF cut-offs, clear differences were seen between the phenotypes of aHUS and C3G at the molecular level. In particular, RV ‘hotspots’ were identified in FH and C3 that could be rationalised on the basis of their importance in protein-protein interactions. For example, FH is involved in both fluid phase and cell surface regulation by (i) being a cofactor for FI, (ii) possessing C3 convertase DAA, and (iii) blocking the formation of the C3 convertase. Despite the FH N-terminal SCR-1/4 and C-terminal SCR-19/20 domains both binding to C3b, required for both fluid phase and cell surface regulation, the C-terminal region of FH is critical for cell surface regulation and is not required for fluid phase regulation (Rodriguez de

[Cordoba et al., 2004](#)). Thus surface-exposed missense RVs in *CFH* that map to individual SCR domains can be correlated with FH function. Other variants predicted to affect FH stability may lead to FH aggregation, making this unable to perform fluid or cell surface phase regulation. In the fluid phase, FH is the only AP complement regulator that has cofactor activity for FI, however at the cell surface, both FH and MCP can act as the FI cofactor. Thus, if the FH C-terminal SCR-19/20 domains are compromised, wild-type MCP may save cell surface regulation. An additional scenario is that if the FH variant is heterozygous, this would affect only half of the FH in plasma, thus altering the resulting phenotype. Overall, the consequence of each RV on FH function can be complex to interpret.

Different FH SCR domains were identified as hotspots in aHUS and C3G:

- (i) For aHUS, the AF analyses confirmed that SCR-20 with 31 RVs and a RV density of 37.2% was a notable missense hotspot ([Figures 4.5A; Appendix IV](#)). SCR-20 is functionally important for FH binding to C3b, C3d, heparin-like oligosaccharides and sialic acid ([Blaum et al., 2015; Rodriguez et al., 2014; Saunders et al., 2006; Schmidt et al., 2008](#)). The occurrence of SCR-20 as a RV hotspot is well explained by the disruption of FH binding to surfaces, leading to host cell damage from excess complement activation caused by unregulated C3b. aHUS missense RVs were also identified in the remaining 19 SCR domains in FH ([Appendix IV](#)). Four of the five FH domains (SCR-2, SCR-5, SCR-8, SCR-12, and SCR-13) do not correspond to known FH binding sites and have only single missense RVs. The distribution ([Figure 4.5A](#)) suggested that the aHUS missense RVs affect mainly FH cell surface binding.
- (ii) In contrast, for C3G, the missense RVs were clustered at the N-terminal C3b binding site (SCR-2/3) ([Figure 4.5A](#)). No missense RVs in C3G were now clustered at cell surface heparin binding sites in SCR-6/7 or SCR-20. The SCR-2/3 domains were identified as C3G hotspots, likely attributed to its binding to MG-2 and MG-6 in C3b ([Figure 4.5A](#)) ([Schmidt et al., 2008; Wu et al., 2009](#)). This different clustering best correlates the C3G variants with dysregulation of the complement AP in the fluid phase (C3, FH) and not at the cell surface (CD46).

Different C3 domains were likewise identified as hotspots in aHUS and C3G:

- (i) For aHUS, the MG-2 and MG-6b domains with the highest RV density ([Appendix IV](#)) were deduced to be RV hotspots ([Figure 4.5B](#)). Both MG domains interact with FH SCR-2 and SCR-3 to enable C3 regulation by FH in both the fluid phase and on the cell surface

(Wu et al., 2009). The disruption of the MG-2 and MG-6b domains would reduce C3 regulation by FH. It is not clear if these two C3 domains also bind MCP thus affecting cell surface regulation further (Schramm et al., 2015). While the TED domain contained 22 RVs, its RV density, which takes into account the number of residues in the domain (300), was not as high as might be expected from its functionally important thioester group and its binding to cell surfaces.

(ii) In distinction, for C3G, too few missense RVs in C3 have been reported for a clear outcome. The RVs were distributed in 11 of its 13 domains with no clustering seen to date.

4.5.5 Utility of the Database

The new Database of Complement Gene Variants enhances the understanding of rare genetic variants in aHUS and C3G for clinical applications. Improvements include the use of AFs, predictive comparisons of wild-type and mutant amino acids, *in silico* analyses using PolyPhen-2 and SIFT, examination of evolution-conserved residues across species, and correlations with functional binding sites. These tools enable clinicians to assess RVs in disease, for example, to investigate which variants within these genes conferred predisposition to aHUS and C3G, and to identify mutational hotspots within these protein structure. This is especially useful for variants of uncertain significance for which no experimental data exists.

Ethnicity data was only recorded for less than 50% of the aHUS and C3G patients in the datasets. The RVs are displayed on the database in comparison with ExAC ethnicity data with a full record of ethnicity. While the disease datasets are incomplete in this regard, the new web-database has the capacity to capture new ethnicity data for aHUS and C3G cases for future AF comparisons. Because the six renal clinics of this study are based in Western Europe and the United States, which are also the source of much of the ExAC dataset, the effect of ethnicity are expected to be minimal on the aHUS and C3G analyses.

Chapter Five

**Two distinct conformations of factor H
regulate discrete complement-binding
functions in the fluid phase and at cell
surfaces**

5.1 Summary

Factor H (FH) is the major regulator of C3b in the alternative pathway of the complement system in immunity. FH comprises 20 short complement regulator (SCR) domains, including eight glycans, and its Tyr402His polymorphism predisposes those who carry it for age-related macular degeneration. To better understand FH complement binding, the solution structures of both the His402 and Tyr402 FH allotypes were studied. In prior work, analytical ultracentrifugation revealed that up to 12% of both FH allotypes self-associate, and this was confirmed by small angle X-ray scattering (SAXS), mass spectrometry and surface plasmon resonance analyses. Here, starting from known structures for the SCR domains and glycans, the SAXS data were fitted using Monte Carlo methods to determine atomistic structures for monomeric FH. The analysis of 29,715 physically realistic but randomised FH conformations resulted in 100 similar best-fit FH structures for each allotype. Two distinct molecular structures resulted that showed either an extended N-terminal domain arrangement with a folded-back C-terminus, or an extended C-terminus and folded-back N-terminus. These two molecular structures are the most accurate to date for glycosylated full-length FH. To clarify FH functional roles in host protection, crystal structures for the FH complexes with C3b and C3dg revealed that the extended N-terminal conformation accounted for C3b fluid phase regulation, the extended C-terminal conformation accounted for C3d binding, and both conformations accounted for bivalent FH binding to anionic glycosaminoglycans on the target cell surface.

5.2 Introduction

The alternative pathway of complement is activated by the spontaneous hydrolysis of C3 into C3u, also known as C3(H₂O). This leads to a positive-feedback amplification of C3 cleavage to form activated C3b that opsonises pathogenic cells. To prevent unwanted C3b-mediated host cell damage, FH regulates complement by acting as a cofactor for FI to cleave C3b (Law & Reid, 1995; Pangburn et al., 1977; Whaley & Ruddy, 1976), competing with FB to interfere with the formation of the C3 convertase C3bBb (Farries et al., 1990), and accelerating the decay of the C3bBb complex (Weiler et al., 1976; Whaley & Ruddy, 1976). These activities occur in the fluid phase and less effectively at the host cell surface (Liszewski et al., 1996).

Disrupted complement regulation is associated with AMD, the most common cause of blindness in the West (Edwards et al., 2005; Hageman et al., 2005; Haines et al., 2005; Klein et al., 2005), and also with aHUS, C3G, and Alzheimer's disease (Osborne et al., 2018a; Saunders et al., 2007; Strohmeyer et al., 2002; Zetterberg et al., 2008). In early stage AMD, sub-retinal pigment epithelial deposits known as drusen develop within Bruch's membrane which is an extracellular matrix layer interposed between the RPE and the choroidal vasculature (Bird, 2010). Drusen contain oxidized lipids, carbohydrates, cellular materials and over 200 proteins including FH and other complement components (Bok, 2005; Crabb et al., 2002; Hageman et al., 2005).

FH is composed of 20 SCR domains of size ~61 residues connected by linkers of lengths 3-8 residues (Figure 5.1) (Lambris & Morikis, 2005). Eight out of nine potential N-glycosylation sites in FH are occupied by biantennary disialylated glycans (Fenaille et al., 2007). Functionally, FH has multiple binding sites for its major ligands C3b, the C3d fragment, C-reactive protein and two host-cell markers heparin, which is an analogue of the natural ligand HS, and sialic acid (Perkins et al., 2012; Perkins et al., 2010a; Perkins et al., 2010b). Each ligand binds to FH weakly with μM K_D values (Perkins et al., 2014) at specific SCR domains summarised in Figure 5.1. For example, FH has two binding sites for heparin at SCR-7 and SCR-20 (Blackmore et al., 1998; Blackmore et al., 1996; Okemefuna et al., 2009; Ormsby et al., 2006; Perkins et al., 2014), and weak independent interactions with heparin at SCR-7 and SCR-20 result in bivalent binding with an overall greater affinity than either site on their own (Khan et al., 2012; Perkins et al., 2012). The dimerization of FH at high concentrations was first reported by X-ray and neutron

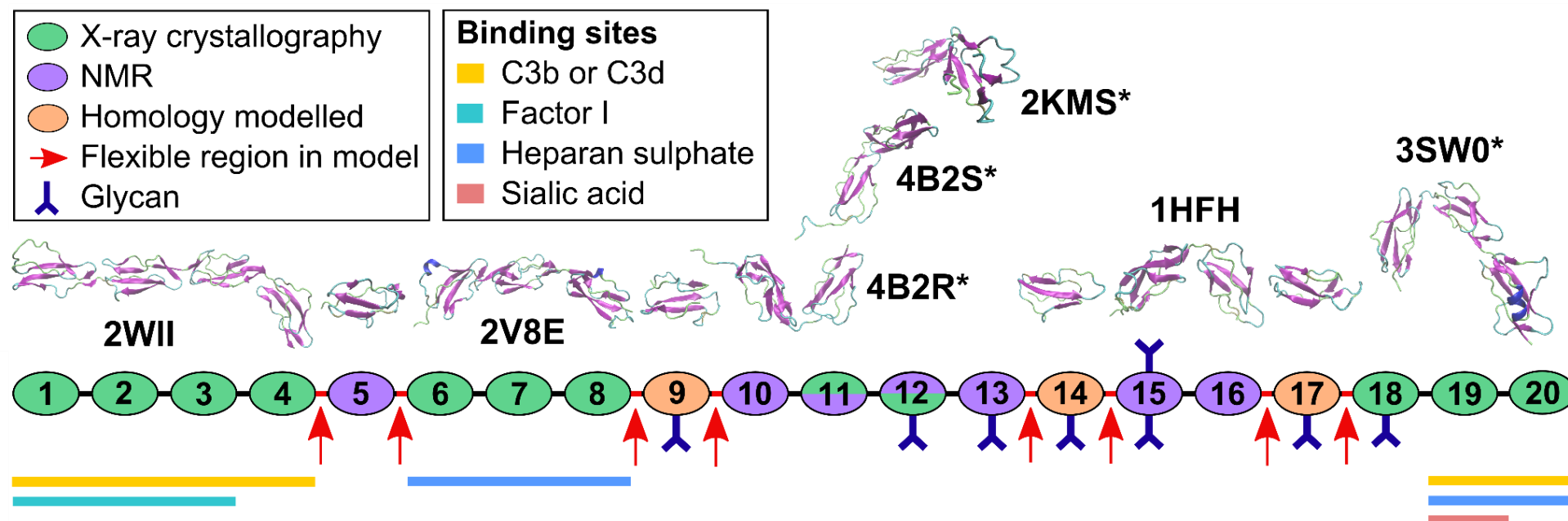


Figure 5.1 Cartoon of the 20 SCR domains in FH. Each SCR domain is represented by an ellipse colour-coded to indicate the starting atomistic structure (green, X-ray crystallography; purple, NMR; orange, homology modelling). Ribbon views of the seven SCR structures used to model FH are shown above the cartoon, together with their PDB codes. The asterisks signify newer SCR structural models that became available after the previous full-length FH models were published. Beneath the cartoon, the FH functional binding sites are denoted by horizontal bars for each of C3b, C3d, factor I, heparan sulphate and sialic acid.

scattering (Perkins et al., 1991). FH also self-associates at each of SCR-6/8 and SCR-16/20 with μM K_D values; the sequential daisy-chaining of these two dimer sites leads to FH oligomers (Fernando et al., 2007; Nan et al., 2008a; Okemefuna et al., 2008; Okemefuna et al., 2009). The formation of dimers and tetramers of FH may also be involved in host cell recognition essential for complement regulation (Pangburn et al., 2009). However it is often presumed that FH is monomeric (DiScipio, 1992).

Genetic variants in FH are associated with disease (Osborne et al., 2018a). In AMD, FH SCR-7 harbours the AMD-risk polymorphism Tyr402His (Edwards et al., 2005; Hageman et al., 2005; Haines et al., 2005; Klein et al., 2005). Individuals homozygous for His402 have a 6-fold increased risk of developing AMD compared to 2.5-fold for His402 heterozygotes (Thakkestian et al., 2006). The presence of His402 weakened FH-heparin binding to both FH SCR-6/8 and full length FH (Herbert et al., 2007a; Skerka et al., 2007), FH binding to C-reactive protein (Laine et al., 2007; Ormsby et al., 2008), and full length FH binding to HS in Bruch's membrane (Langford-Smith et al., 2014). For the development of AMD, age-related changes in glycosaminoglycan structures such as a decrease in the sulfation of HS, together with the presence of FH His402, have been proposed to reduce FH binding over time. At host cell surfaces, FH His402 showed weaker SCR-7 binding to eye tissue-specific HS only (Clark et al., 2013). In opposition to these observations of weaker HS binding for the His402 allotype, His402 in SCR-6/8 was seen to be bound to the highly-sulphated glycosaminoglycan analogue sucrose octasulfate using X-ray crystallography while no such crystals were reported for the Tyr402 allotype (Prosser et al., 2007b). Although the statistical association of Tyr402His with AMD is one of the strongest, many other complement variants in *CFH* and other complement genes associate also with AMD, suggesting that AMD is caused by several different molecular mechanisms. For example, at the other heparin binding site on FH, SCR-20 harbours the AMD-risk RV Arg1210Cys (Raychaudhuri et al., 2011). AMD is associated with another 12 RVs in FH in its signal peptide or in SCR-1, SCR-3/5, SCR-8, SCR-16 or SCR-18 (Geerlings et al., 2017). AMD is also associated with the protective polymorphism Val62Ile in SCR-1 (Hageman et al., 2005; Hocking et al., 2008), common synonymous and non-coding variants in a region overlapping the *CFH* gene (Herbert et al., 2007b; Li et al., 2006; Maller et al., 2006), and genetic variants in complement C3, FI, FB and C9 (Geerlings et al., 2017; Hecker et al., 2010).

Additional insights into molecular mechanisms for AMD may be gained by studying both the self-association and overall solution structure of FH in the context of the Tyr402His polymorphism. To date, full-length FH has proven too large, flexible and glycosylated for high resolution structural determination by crystallography or NMR. Preliminary molecular modelling of SAXS curves revealed that full-length FH in solution has a folded-back SCR structure that is affected by ionic strength (Aslam & Perkins, 2001; Nan et al., 2010; Okemefuna et al., 2009). This preliminary FH modelling was based on just 11-14 high resolution SCR structures and an initial molecular dynamics approach to model the inter-SCR linkers (Nan et al., 2010; Okemefuna et al., 2009). In prior work, homozygous FH Tyr402 and His402 were submitted to analytical ultracentrifugation (AUC), SAXS, mass spectrometry and surface plasmon resonance studies, which found that both FH allotypes self-associate to form oligomers (Osborne et al., 2018b). Here, in order to examine the effect of the Tyr402His polymorphism on the solution structure of FH, SAXS curve extrapolation to zero concentration showed that both FH allotypes exhibited similar overall structures. The SAXS curves for both FH allotypes were modelled by starting from 17 high resolution SCR structures (Figure 5.1; Table 5.1) (Hocking et al., 2008; Okemefuna et al., 2009; Schmidt et al., 2010; Wu et al., 2009) and utilising a much improved Monte Carlo atomistic modelling method for the full-length FH structure, including the eight FH glycan chains (Perkins et al., 2016). The resulting SAXS best-fit structural ensembles corresponded to the most accurate FH models determined to date. These ensembles revealed two different extended and folded-back FH domain arrangements that favour two different modes of FH binding to either C3b or C3d. Both arrangements accounted for the bivalent binding of FH to HS-coated cell surfaces. The outcome of two different FH conformations provides new insights into the way in which FH regulates complement C3b activity.

5.3 Methods

5.3.1 Homology modelling of SCR-9, SCR-14 and SCR-17

The first scattering modelling of the 20 SCR domains in FH used NMR structures for three SCR domains and homology models for the remaining 17 SCR domains (Figure 5.1) (Aslam & Perkins, 2001). The second modelling used NMR and crystal structures for 11-14 SCR domains, and homology models for the remaining 9 SCR domains (Nan et al., 2010; Okemefuna et al., 2009). Here, a combination of MODELLER v9.14

Table 5.1 Sources of molecular SCR structures in FH.

FH domain(s)	PDB code	Reference	Method	Used in previous FH model
SCR-1/4	2WII	(Wu et al., 2009)	X-ray crystallography	(Nan et al., 2010)
SCR-1/3	2RLP/2RLQ	(Hocking et al., 2008)	NMR	(Okemefuna et al., 2009).
SCR-4		(Saunders et al., 2006)	Homology	(Okemefuna et al., 2009).
SCR-5		(Barlow et al., 1992)	NMR	(Okemefuna et al., 2009); (Nan et al., 2010)
SCR-6/8	2V8E	(Prosser et al., 2007a)	X-ray crystallography	(Okemefuna et al., 2009); (Nan et al., 2010)
SCR-9 b		(Saunders et al., 2006)	Homology	(Okemefuna et al., 2009); (Nan et al., 2010)
SCR-10/11 ^a	4B2R	(Makou et al., 2012)	NMR	-
SCR-10/13		(Saunders et al., 2006)	Homology	(Okemefuna et al., 2009).
SCR-11/12 ^a	4B2S	(Makou et al., 2012)	X-ray crystallography	-
SCR-12/13	2KMS	(Schmidt et al., 2010)	NMR	(Nan et al., 2010)
SCR-14 ^b		(Saunders et al., 2006)	Homology	(Okemefuna et al., 2009); (Nan et al., 2010)
SCR-15/16	1HFH	(Barlow et al., 1993)	NMR	(Okemefuna et al., 2009); (Nan et al., 2010)
SCR-17 ^b		(Saunders et al., 2006)	Homology	(Okemefuna et al., 2009); (Nan et al., 2010)
SCR-18		(Saunders et al., 2006)	Homology	(Okemefuna et al., 2009).
SCR-18/20 ^a	3SW0	(Morgan et al., 2012)	X-ray crystallography	-
SCR-19/20	2G7I	(Jokiranta et al., 2006)	X-ray crystallography	(Okemefuna et al., 2009); (Nan et al., 2010)

^a Newer SCR structural models published after the previous full-length FH models were published.

^b Only homology models are available for this (Figure 5.1).

(Sali & Blundell, 1993) and monomer Monte Carlo (SASSIE-web) (Curtis et al., 2012) was used to build a starting FH model from NMR and crystal structures for 17 SCR domains and three improved SCR homology models for SCR-9, SCR-14 and SCR-17 (Tables 5.1 and 5.2). For the homology modelling, a template SCR structure without linker residues was identified using the basic local alignment search tool to search the PDB server (BLAST-PDB) (Altschul et al., 1990) (Table 5.2), and its sequence was aligned to the template sequence using Clustal Omega multiple sequence alignment (Sievers et al., 2011). By this, low sequence identity was mostly confined to the loop regions, and the secondary structure was conserved. These loop regions were modelled using the loop optimisation protocol in MODELLER. The best SCR model was selected from a final dataset of 100 generated models using the normalised discrete optimised protein energy score (Shen & Sali, 2006). For the SCR-9, SCR-14 and SCR-17 homology models and the NMR-based SCR-5 model (Barlow et al., 1992), strained high energy bonds were relaxed to give physically stable models by energy minimisation (500,000 steps) and molecular dynamics simulations (1,000,000 steps) in the simulation engine NAMD v2.9 (Phillips et al., 2005). The PDB Reader input generator tool (Jo et al., 2014; Jo et al., 2008) converted the PDB file into CHARMM readable format, generated a PSF including the disulphide bonds, and provided the CHARMM36 force-field files (Brooks et al., 2009; Lee et al., 2016). The quality of the secondary structure was verified using Ramachandran plots on the RAMPAGE server (Lovell et al., 2003), the DSSP program (Joosten et al., 2011; Kabsch & Sander, 1983) and visual comparison with other SCR domains. As required, the Ramachandran plots showed a minimal number of outliers in these four SCR models.

5.3.2 Building of initial FH model with eight glycans

In order to create the initial full-length FH structure (Figure 5.1), the 11 SCR fragment models (SCR-1/4, SCR-5, SCR-6/8(His402), SCR-9, SCR-10/11, SCR-11/12, SCR-12/13, SCR-14, SCR-15/16, SCR-17, SCR-18/20) were aligned to the FH sequence sequentially using MODELLER. Each of SCR-1/5, SCR-6/13, SCR-14/16 and SCR-17/20 were assembled first, before their simultaneous alignment to the full-length FH sequence. The best FH model was selected based on the normalised discrete optimised protein energy score, then refined in NAMD to give the final FH initial structure. FH was numbered from the N-terminus of the 18-residue signal peptide to follow Human Genome Variation Society convention. Eight biantennary disialylated glycans

Table 5.2 Homology modelling of the SCR-9, SCR-14, and SCR-17 domains.

Target domain	Template domain	PDB code	Residues	Maximum score	Total	Query cover	E value ^a	Residue identities
SCR-9	SCR-4	2UWN	148-186	40.1	41.2	68%	0.0004	44%
SCR-14	SCR-15	3ZD1	5-59	37.5	37.5	100%	0.002	36%
SCR-17	SCR-18	3SW0	3-59	42.2	77.6	100%	0.00007	37%

^a Expect value. This is a parameter that describes the number of hits one can expect to see by chance.

(Fenaille et al., 2007) were added to Asn511 in SCR-9, Asn700 in SCR-12, Asn784 in SCR-13, Asn804 in SCR-14, Asn864 and Asn892 in SCR-15, Asn1011 in SCR-17 and Asn1077 in SCR-19 (Figure 5.1). For each glycan-Asn PDB files, GlycanReader was used to generate the CHARMM force field and PSF inputs for energy minimisation in NAMD. The resulting energy-minimised glycans were positioned onto FH by superimposing the Asn residues in PyMol, then deleting the Asn residue in the glycan PDB. Once all eight glycans had been added to FH and accepted by GlycanReader, bash scripts were used to convert the nomenclature and numbering of the glycan and protein atoms to the format required for the Torsion Angle Monte Carlo module in SASSIE-web. In this, the glycans were moved with their attached protein segment as a single entity and not varied independently.

5.3.3 Building of FH model library using Monte Carlo

The model library of physically realistic FH structural conformations was generated by subjecting the inter-SCR linkers to the Torsion Angle Monte Carlo module in SASSIE-web (Zhang et al., 2017). Eleven of the 19 linkers were defined by high resolution structures and were therefore held fixed. The remaining eight linkers at SCR-4/5, SCR-5/6, SCR-8/9, SCR-9/10, SCR-13/14, SCR-14/15, SCR-16/17, and SCR-17/18 were moved (arrowed, Figure 5.1). For each of these linker residues, the backbone ϕ and ψ torsion angles were varied in steps of up to either 30° or 180° , except for the torsion angles involving the conserved Cys residues at the start and end of each linker. In order to maximise the sampling, four simulations were run as follows: (i) 200,000 steps with up to 30° moves using the FH initial model, (ii) 200,000 steps with up to 180° moves using the FH initial model, (iii) 100,000 steps with up to 30° moves using a FH model from simulations (i, ii) with its CT bent inwards, and (iv) 10,000 steps with up to 30° moves using the best-fit FH model from simulations (i-iii). In a Monte Carlo simulation, many moves result in structures with steric clashes, and were discarded as physically unrealistic. In the present case, the 510,000 attempted moves resulted in 15,136 (8%), 3720 (1%), 10,316 (10%) and 543 (5%) physically-realistic, acceptable models respectively. These were combined into a library of 29,715 models for SAXS curve fitting.

5.3.4 Fitting FH models to experimental scattering data

Prior to the work carried out for this PhD thesis, experimental SAXS data for two FH Tyr402 and two FH His402 homozygous allotypes were acquired at five concentrations between 0.4 mg/ml (2.6 μ M) and 3.3 mg/ml (21.4 μ M) (Table 5.3) (Osborne et al., 2018b). Before these SAXS experiments, the Tyr402His and Ile62Val polymorphisms were identified by direct DNA sequencing of their PCR product (Osborne et al., 2018b). Here, in this PhD thesis chapter, a theoretical scattering curve was generated from each of the 29,715 FH models by using SasCalc. SasCalc calculates the scattering curve using an exact all-atom expression for the scattering intensity in which the orientations of the Q vectors are taken from a quasi-uniform spherical grid generated by the golden ratio (Watson & Curtis, 2013). For each of the Tyr402 and His402 allotypes, the experimental scattering curve with 221 data points was extrapolated to zero concentration to eliminate the contribution of self-association. These were compared to the theoretical curves using the R -factor:

$$R = \sum_{Q_i} \frac{\|I_{\text{expt}}(Q_i) - I_{\text{theor}}(Q_i)\|}{\|I_{\text{expt}}(Q_i)\|} \times 100 \quad (5.1)$$

where Q_i is the Q value of the i -th data point, $I_{\text{exp}}(Q_i)$ is the experimental scattering intensity and $I_{\text{model}}(Q_i)$ is the theoretical modelled scattering intensity (Wright & Perkins, 2015). For SasCalc, the lowest Q values before the Guinier R_G region in the extrapolated scattering curves were interpolated to zero Q using MATLAB. After interpolation, the original 221 $I(Q)$ values between Q of 0.0-1.5 nm^{-1} were retained, to define the Q spacing for SasCalc. The R -factor vs R_G graphs were similar for the extrapolated and 0.4 mg/ml curves, while those from curve fitting at 0.7, 1.1 and 2.2 mg/ml gave worse R -factors (data not shown), as expected. For each of the four Tyr402 and His402 curves, the 29,715 models were filtered on both R_G and R -factor. Models were accepted if their R_G values were within $\pm 5\%$ of the experimental R_G values (Table 5.3), and their R -factor was $\leq 5\%$. The best-fit 100 models were identified by ranking the filtered models by their R -factors. The Tyr402His polymorphism had no effect on the curve fits, leading to an R -factor difference of only 0.0003%. In order to analyse the flexibility of FH, normalised Kratky

Table 5.3 Experimental X-ray and analytical ultracentrifugation data for FH Tyr402/Val62 and His402/Val62 and their modelling fits

	Filter	Models	R_G (nm)	R_{XS-1} (nm)	R_{XS-2} (nm)	D_{max} (nm)	R factor (%)	$s^0_{20,w}$ (S)
<i>Experimental data</i> (Osborne et al., 2018b)								
FH Tyr402/Val62			7.39 ± 0.25	2.21 ± 0.06	1.77 ± 0.01	25		5.66 ± 0.08
			7.35 ± 0.13	2.27 ± 0.06	1.77 ± 0.01			
			7.61 ± 0.01^a					
FH His402/Val62			7.77 ± 0.27	2.02 ± 0.06	1.76 ± 0.01	25		5.69 ± 0.02
			7.22 ± 0.15	2.15 ± 0.06	1.75 ± 0.01			
			7.54 ± 0.01^a					
<i>Atomistic Modelling</i>								
Library of 29,715 FH models	None	29,715	5.61 - 11.08	n.a	n.a	n.a	2.8 - 25.3	n.a.
FH Tyr402/Val62 (sample 1)	R_G and R -factor	1,240	7.02 – 7.75	n.a	n.a	n.a.	2.3 – 4.9	n.a.
	Best fit	100	7.02 – 7.57	2.34 – 2.98	1.85 – 2.10	n.a.	2.3 – 3.2	5.35 – 5.77
	PCA group 2 (NT out)	49	7.34 – 7.57	2.34 – 2.62	1.95 – 2.10	n.a.	2.7 – 3.2	5.38 – 5.77
	centroid	1	7.41	2.50	2.03	n.a.	3.1	5.52
FH Tyr402/Val62 ^b (sample 2; data not shown)	R_G and R -factor	2,678	6.98 – 7.72	n.a	n.a	n.a.	1.9 – 4.9	n.a.
	Best fit	100	7.03 – 7.47	2.71 – 3.05	1.81 – 2.04	n.a.	1.9 – 2.7	5.33 – 5.70
	PCA group 14 (NT in)	48	7.11 – 7.43	2.71 – 2.98	1.81 – 1.99	n.a.	1.9 – 2.7	5.36 – 5.70
	centroid	1	7.32	2.84	1.95	n.a.	2.2	5.49
FH His402/Val62 (sample 1)	R_G and R -factor	1,219	7.38 – 8.16	n.a	n.a	n.a.	2.3 – 4.9	n.a.
	Best fit	100	7.39 – 8.13	2.32 – 3.15	1.48 – 1.97	n.a.	2.4 – 3.6	5.21 – 5.61
	PCA group 5 (NT in)	79	7.39 – 7.95	2.67 – 3.07	1.64 – 1.94	n.a.	2.5 – 3.6	5.21 – 5.55
	centroid	1	7.48	2.92	1.87	n.a.	3.2	5.44
FH His402/Val62 (sample 2; data not shown)	R_G and R -factor	310	6.86 – 7.58	n.a	n.a	n.a.	2.7 – 4.9	n.a.
	Best fit	100	6.86 – 7.57	2.09 – 2.93	1.82 – 2.10	n.a.	2.7 – 4.0	5.38 – 5.88
	PCA group 19 (NT out)	84	6.86 – 7.57	2.28 – 2.82	1.89 – 2.10	n.a.	2.7 – 4.0	5.38 – 5.78
	centroid	1	7.40	2.48	2.03	n.a.	3.8	5.51

^a Both the R_G and R_{XS} values correspond to the SAXS curve extrapolated to zero concentration. The first R_G value corresponds to the Guinier fits; the second corresponds to the $P(r)$ analyses.

^b In this instance Ile62Val was heterozygous; the other three were homozygous for Val62.

n.a, not available

analyses of $(Q/R_G)^2 \cdot I(Q)/I(0)$ vs Q/R_G were calculated for the four FH experimental curves (Receveur-Brechot & Durand, 2012).

5.3.5 Parameterisation of the FH domain arrangement

The SCR domain arrangement in the 29,715 FH models was parametrised using the separations between the FH centre of mass (COM) and the N-terminal (NT) and C-terminal (CT) α -carbon atoms. Separations were calculated using Tcl scripting within Visual Molecular Dynamics (VMD) software (Humphrey et al., 1996), and their frequencies were displayed as histograms using R (R Core Team, 2013) for statistical analyses. In order to assess the normality of the separation distributions, Quantile-Quantile (Q-Q) plots and two-tailed Kolmogorov-Smirnov tests were employed using the `ks.test` statistical function in R (Grant et al., 2006; R Core Team, 2013). Here, the null hypothesis stated that the distribution was consistent with the normal distribution, and a Bonferroni-corrected significance level of 0.05 divided by 2 (0.025) was set. The Kolmogorov-Smirnov test calculated the maximum vertical deviation between the two curves as the test statistic D and the probability of D occurring due to chance (p -value). In order to compare the distributions for all 29,715 models, the 310-2,678 filtered models, and the best-fit 100 models of the two allotypes, their distributions were converted to cumulative density frequencies and compared using the two-tailed Kolmogorov-Smirnov test in R with the `ks.test` statistical function. Here, for each comparison, the null hypothesis stated that the distributions were the same and a Bonferroni-corrected significance level of 0.05 divided by 12 (0.00417) was set. In order to visualise the resulting conformations after applying the R_G and R -factor filters, the filtered models were superimposed with the reference FH model using SCR-10/13, and visualised as density plots using SASSIE-web.

5.3.6 Principal component analyses (PCA)

PCA provided by the Bio3d package in R (Grant et al., 2006) identified four major conformational states in the best-fit 100 models. For both allotypes, PCA satisfactorily accounted for >80% of the variance between the models. The average FH structure for each PCA group was identified using a centroid model computed using R. The SCR domain arrangement of each PCA group was identified by a density plot using SASSIE-web. In order to assess the conformational differences between the two allotypes, the

frequencies of the two major SCR domain arrangements were compared by using the two-tailed Fisher's test for a 2×2 contingency table in GraphPad QuickCalcs (<https://www.graphpad.com/quickcalcs/>). Here, the null hypothesis stated that the proportion of FH models with an inwardly-bent NT and extended CT, compared to an extended NT and inwardly-bent CT, were the same. The significance level was set as 0.05 and the probability of the difference occurring due to chance (*p*-value) was computed. To see whether these two conformations could be distinguished by their SAXS curves, the difference in scattering intensity between their best fitting centroid model curves was calculated and visualised using R. The FH models were assessed by superimposing the centroid model for each of the largest ensembles with crystal structures for the FH-ligand complexes of SCR-1/4 with C3b and SCR-19/20 with C3d using PyMol (PDB codes: 2wii and 5nbq) (Kolodziejczyk et al., 2017; Wu et al., 2009).

5.3.7 Theoretical sedimentation co-efficient values

Prior to the work carried out for this PhD thesis, AUC sedimentation velocity experiments were conducted on five homozygous FH Tyr402 and five homozygous FH His402/Val62 samples (Osborne et al., 2018b). Here, in this PhD thesis, for the AUC modelling, the theoretical $s_{20,w}^0$ values for the FH models were calculated directly from the atomic coordinates with the default value of 0.31 nm for the atomic element radius for all atoms to represent the hydration shell by using the HYDROPRO shell modelling program (Garcia-de la Torre et al., 2000). The molecular weight and partial specific volume for FH used followed that used previously (Okemefuna et al., 2009; Ortega et al., 2011). FH glycosylation was considered by using the compiled partial specific volumes for saccharides (Perkins, 1986).

5.3.8 Links to web servers and tools

BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>),
 Clustal Omega (<http://www.ebi.ac.uk/Tools/msa/clustalo/>),
 SASSIE-web (<https://sassie-web.chem.utk.edu/sassie2/>),
 CHARMM-GUI Glycan Reader (<http://charmm-gui.org/?doc=input/glycan>),
 CHARMM-GUI PDB Reader (<http://www.charmm-gui.org/?doc=input/pdbreader>),
 RAMPAGE (<http://mordred.bioc.cam.ac.uk/~rapper/rampage.php>),
 DSSP (<http://swift.cmbi.ru.nl/gv/dssp/>),

GraphPad QuickCalcs (<https://www.graphpad.com/quickcalcs/contingency1/>).

5.4 Results

5.4.1 FH models for the Tyr402 and His402 allotypes

The atomistic modelling of the two FH Tyr402 and two His402 scattering curves was initiated using 17 high-resolution SCR structures, three SCR homology models and eight glycan chains to construct an energy-minimised initial FH model (Figure 5.1, Tables 5.1 and 5.2; Section 5.3). By varying the torsion angles at eight inter-SCR linkers, 510,000 FH models were created in four Monte Carlo simulations, from which 29,715 models were retained for analysis as being physically realistic without steric clashes. Comparison of the four extrapolated experimental scattering curves with the 29,715 theoretical curves gave an R -factor vs. R_G distribution with a minimum close to the experimental extrapolated R_G value (Figure 5.2). An R_G filter of $\pm 5\%$ of the experimental value and an R -factor filter of $>5\%$ gave 1,240/2,678 and 1,219/310 best-fit models for the two Tyr402 and two His402 allotypes respectively, these being 1-9% of the 29,715 trial models (blue/red, Figure 5.2). The best-fit 100 FH models were selected on the basis of the lowest R -factors (orange, Figure 5.2) of 2.3-3.2% and 2.3-3.6%, respectively (Table 5.3).

The $s_{20,w}^o$ values of the best-fit 100 FH models were calculated using HYDROPRO to be 5.35-5.77 S and 5.21-5.61, and for the centroid models to be 5.49-5.52 S and 5.44-5.51 S, both for FH Tyr402 and FH His402 respectively (Table 5.3). These values agreed well with the previously obtained experimental values of 5.66-5.69 S (Table 5.3) (Osborne et al., 2018b), and confirmed the outcome of the atomistic modelling, given that the mean difference between the modelled and experimental values should be ± 0.21 S for related macromolecules including antibodies (Perkins et al., 2009). Like antibodies which are also considered to be flexible molecules, HYDROPRO was effective in calculating the $s_{20,w}^o$ values, this being attributed to the use of the averaged best-fit solution structures. The calculations were much improved to those of the 2009 FH modelling which gave 4.91-5.01 S from older scattering models that in retrospect were too elongated (Okemefuna et al., 2009).

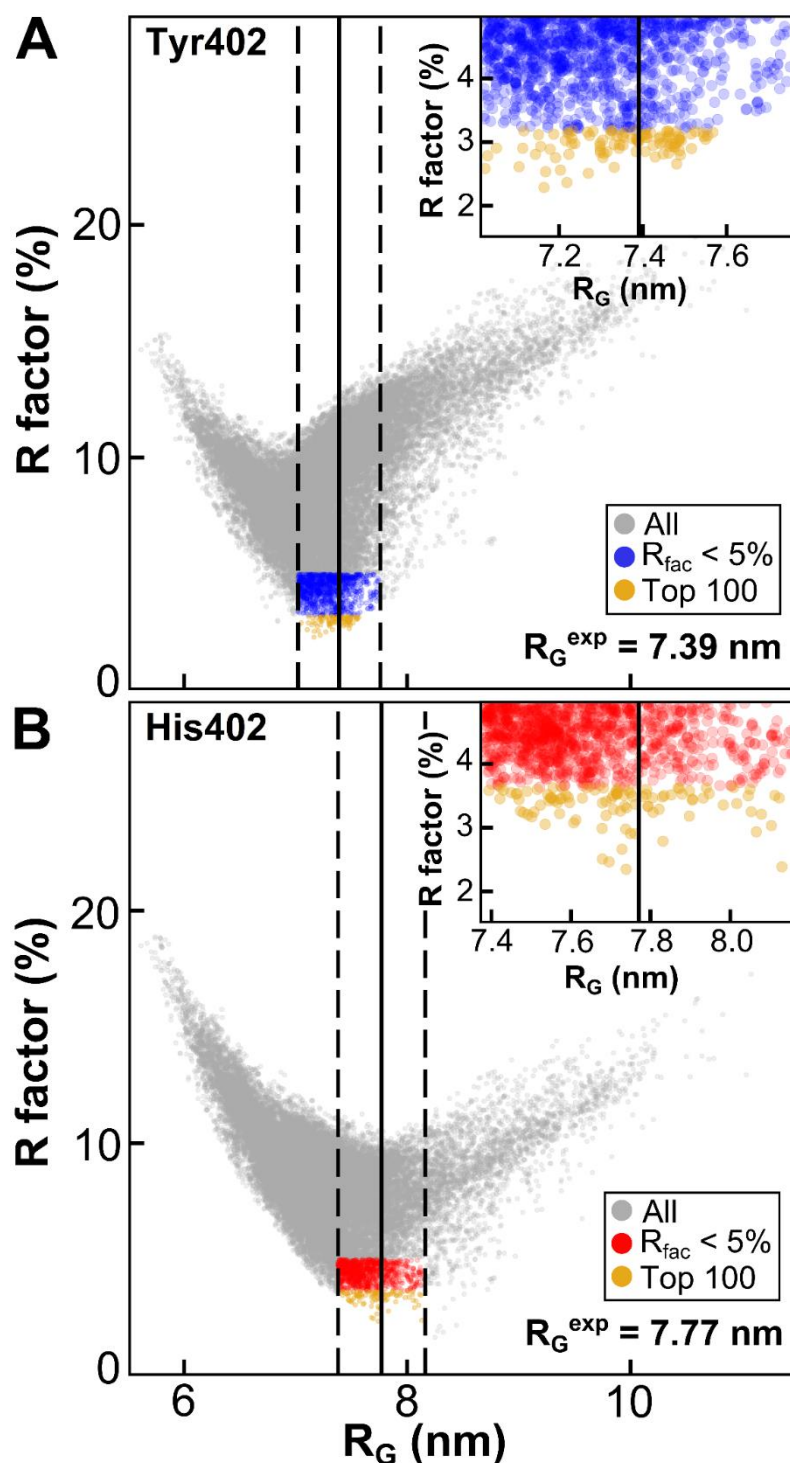


Figure 5.2 Atomistic modelling searches for the FH solution structure. The 29,715 FH models were fitted to one experimental scattering curve for each of *A*, homozygous FH Tyr402/Val62 and *B*, FH His402/Val62, each extrapolated to zero concentration. Grey corresponds to all FH models. The blue/red subsets corresponds to FH models that passed two filters, namely R_G within $\pm 5\%$ of the experimental R_G , and an R -factor of $< 5\%$. Orange corresponds to the 100 best-fit FH models (see inset). The vertical black lines represent the experimental R_G for FH Tyr402 (7.39 nm) and FH His402 (7.77 nm). The dashed lines represent the $\pm 5\%$ upper and lower boundaries of these R_G values.

5.4.2 The NT-COM and CT-COM separation distributions

The best-fit SCR arrangements within the 29,715 trial FH models were shown to be bimodal using histograms of their N-terminal α -carbon (NT) and C-terminal α -carbon (CT) to centre of mass (COM) separations (Figure 5.3). Longer separations corresponded to extended SCR arrangements in FH; shorter separations corresponded to a folded-back bent one.

(i) For all 29,715 models, the NT-COM and CT-COM separations corresponded to similar ranges of 0.5-21.8 nm and 0.8-22.4 nm respectively (Figure 5.3A,D). However, the most populated NT-COM separations were short ones (3-5 nm) that tailed off after >16 nm. In distinction, the most populated CT-COM separations were long ones (14-16 nm) followed by short ones (~ 5 nm) (Figure 5.3A,D). The NT-COM separations were right skewed with a skewness coefficient of 0.53, whereas the CT-COM separations were left skewed and bimodal with a skewness coefficient of -0.55. Overall, the separation distributions were uneven and not normally distributed. This unexpected outcome was attributed to the seven glycans in SCR-12/18 that perturbed the random generation of physically-realistic FH structures (Figure 5.1) and the asymmetric distribution of the longest inter-SCR linkers in FH (Okemefuna et al., 2009).

(ii) For the 1219-1240 filtered FH models, both the NT-COM and CT-COM distance distributions corresponded to small subsets of those for all 29,715 models (Figure 5.3B,E). Density plots showed that these models occupied limited conformations when compared to all 29,715 models (insets, Figure 5.3B,E). The separations for both allotypes were bimodal, being left-skewed for the NT-COM distances, and right-skewed for the CT-COM distances. Overall, no differences were seen between the two allotypes.

(iii) For the two sets of 100 best-fit FH models, the two NT-COM and CT-COM distance distributions were again largely bimodal distributions (Figure 5.3C,F). Similar bimodal results were obtained for the fits with the other pair of FH Tyr402 and FH His402 experimental curves (distributions not shown).

For the filtered and top 100 models, statistical analyses were performed to confirm that the separation distance distributions (Figure 5.3) were unaffected by sampling bias. The best-fit 100 models were not necessarily a converged subset of the filtered models as

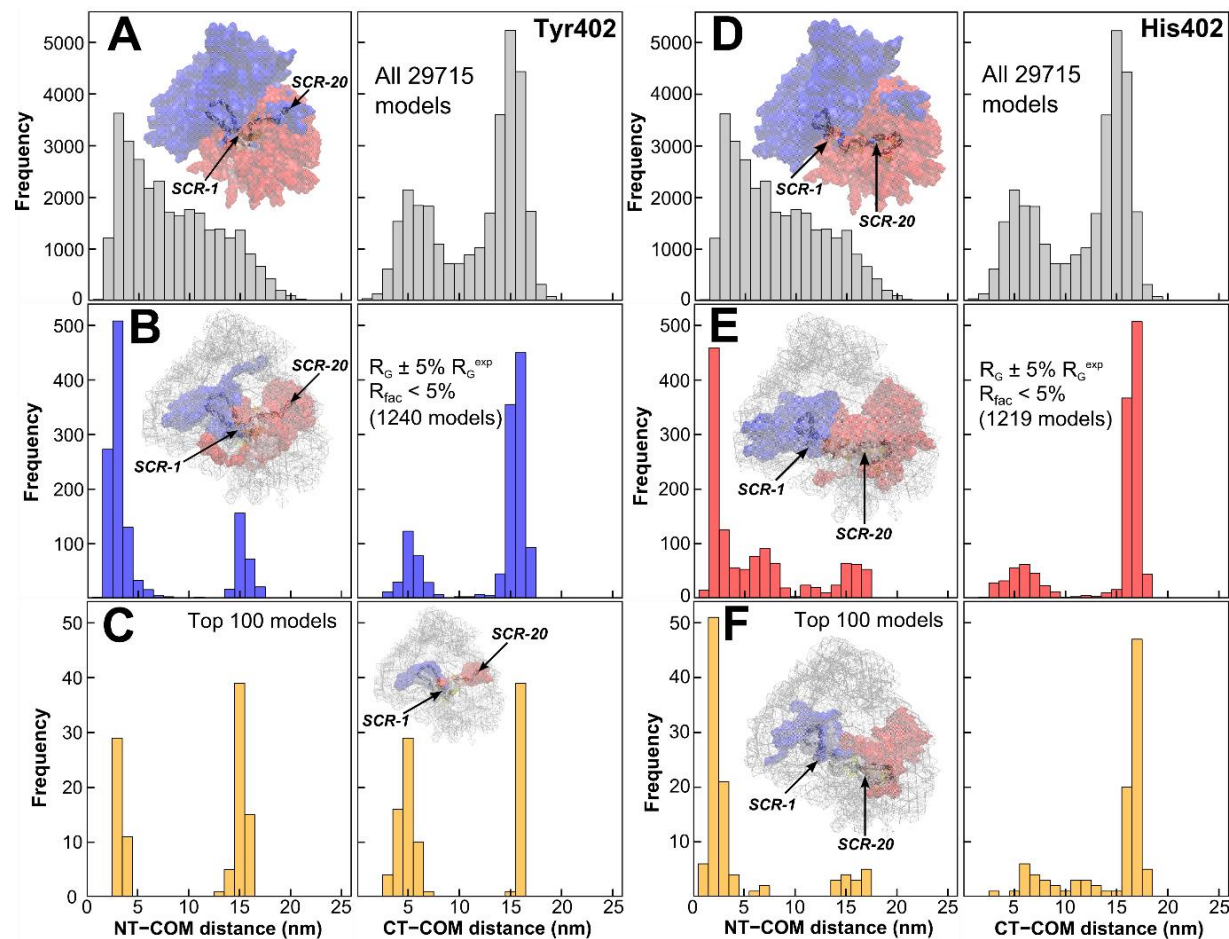


Figure 5.3 The centre-of-mass separation frequencies in the FH Tyr402 and FH His402 models. A, D, The top two panels (grey) show the separations (NT-COM, CT-COM) between the N-terminal α -carbon (NT) and C-terminal α -carbon (CT) to the centre of mass (COM) for all 29,715 FH models. B, The separations are shown for the filtered 1,240 and 1,219 FH models (see Figure 5.2). C, F, The separations are shown for the 100 best-fit FH models. For each set of FH models, the density plot shows the N-terminal half in blue, the C-terminal half in red, and the best-fit model as a cartoon.

seen in other modelling (Fung et al., 2016). A non-parametric (i.e. no underlying assumption) Kolmogorov-Smirnov test (not shown) was used to indicate that the best-fit distances for NT-COM and CT-COM were either short or long, and were independent of any sampling bias observed in all 29,715 models. This suggested that FH existed in two conformations. In order to check that the distance parameters for the filtered models were independent of sampling bias, each distance distribution (Figure 5.3) was converted to a cumulative density distribution (Figure 5.4A-D). For both allotypes, the cumulative distributions for each of the filtered and best-fit 100 models (red, blue, orange) were significantly different from that (in grey) for all 29,715 models (each $p < 2.2 \times 10^{-16}$; two-tailed Kolmogorov-Smirnov test; arrowed as A, B, C, D and G, H, I, J in Figure 5.4A-D). This was as expected because only a small subset of the starting models would fit the scattering curves. This also indicated that the results for the filtered and best-fit 100 models were independent of any sampling bias seen in all 29,715 models. A significant difference was also seen between the filtered and top 100 models in three of the four cases ($p = 2.5 \times 10^{-12}$ or $p = 1.31 \times 10^{-7}$; two-tailed Kolmogorov-Smirnov test; excepting that for CT-COM distance for His402 ($p = 0.327$); E, F, K, L in Figure 5.4A-D). Overall, because the best-fit 100 models were not necessarily a converged subset of the filtered models as seen in other modelling (Fung et al., 2016), it was concluded that the best-fit distances for NT-COM and CT-COM were either short or long, suggesting that FH existed in two conformations.

The full FH SCR domain arrangement was examined using pairs of NT-COM and CT-COM distances. The 29,715 separations were non-randomly distributed between 1-25 nm, showing that a broad range of FH conformations had been generated as desired (Figure 5.4E,F). The most sampled best-fit NT-COM and CT-COM separations occurred at 1-10 nm and 12-17 nm respectively at the top left hand corner. The second most sampled best-fit separations occurred as a broad cluster at 5-17 nm and 2-8 nm respectively. For the filtered Tyr402 models, two structural regions were visible as two dense clusters in blue, in which the top 100 Tyr402 models were shown in orange (Figure 5.4E). For the filtered His402 models, the same two clusters were seen in red and orange respectively (Figure 5.4F). In contrast, FH models with long NT-COM and CT-COM separations (>15 nm) showed large radii of gyration and poor fits with R-factors $>5\%$ (grey, Figure 5.4E,F). This analysis showed that FH either has a bent-inwards N-terminal region and an extended C-terminal region, or vice versa.

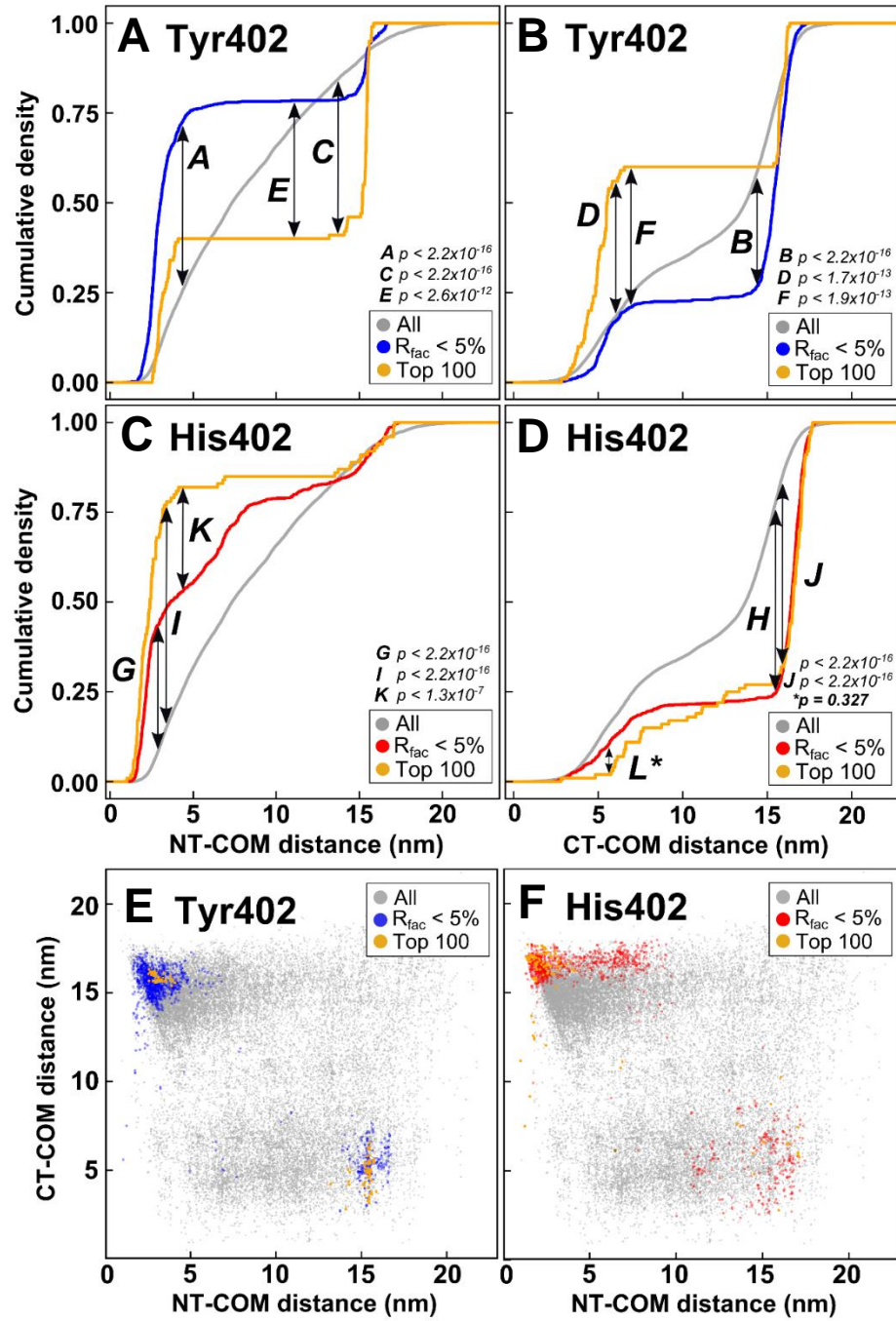


Figure 5.4 The separation densities in the FH Tyr402 and FH His402 models. For each of the Tyr402 (A, B) and His402 (C, D) allotypes, the cumulative densities of the individual NT-COM and CT-COM separations are shown for all 29,715 FH models (grey lines), the filtered FH models and the 100 best-fit FH models (orange lines). The double-headed arrows labelled A-L) indicate the maximum vertical deviation of the densities compared with each other using a combination of quantile-quantile plots and the Kolmogorov-Smirnov statistical test, of which L denoted by * is non-significant. (E,F), The joint NT-COM and CT-COM separations for the FH Tyr402 and FH His402 models are compared with each other. Blue/red corresponds to the filtered FH models and orange corresponds to the 100 best-fit FH.

5.4.3 Best-fit conformations for FH Tyr402 and His402

After superimposition of the top 100 Tyr402 and His402 FH structures, visual inspection showed multiple conformers. To understand these conformers, they were clustered into conformational families using principal component analysis (PCA) (Figure 5.5) (Hui et al., 2015). PCA determines the correlated motions of protein residues as linearly uncorrelated variables termed principal components (David & Jacobs, 2014). These “essential motions” are extracted from a covariance matrix of the atomic coordinates of the frames in the trajectory. The eigenvectors of this matrix each have an associated eigenvalue that characterises a mode of the motion, or variance.

The PCA analyses confirmed that two families of FH structures with either a bent-inwards N-terminal region and an extended C-terminal region, or vice versa, were seen. Four PCA groups accounted for 98.5% and 86.9% of the variances in the top 100 models of each of Tyr402 and His402 respectively (Figure 5.5C,G). For each PCA group, the combined NT-COM and CT-COM separation distances were plotted (Figure 5.5D,H). The conformational density of each group was represented as either blue (NT) or red (CT) density, with the centroid FH model depicted as a black cartoon (Figure 5.6). The 60 Tyr402 models in PCA groups 2, 3 and 4 corresponded to an extended N-terminus and a bent inwards C-terminus (red, green and blue, Figure 5.5D; Figure 5.6B-D). The 40 Tyr402 models in PCA group 1 corresponded to a bent inwards N-terminus and an extended C-terminus (black, Figure 5.5D; Figure 5.6A). The 79 His402 models in PCA group 5 corresponded to a bent inwards N-terminus and an extended C-terminus (black, Figure 5.5H; Figure 5.6E). The 21 His402 models in PCA groups 6, 7 and 8 corresponded to a bent inwards C-terminus and an extended N-terminus (red, green and blue, Figure 5.5H; Figure 5.6F-H).

Each of the PCA groups 1-8 gave FH structures whose theoretical scattering curves $I(Q)$ fitted well with the experimental Tyr402 or His402 curves with low R-factors between 2.4%-3.8% (Figure 5.6). The visual agreement was good out to at least $Q = 1.0 \text{ nm}^{-1}$. The centroid models with an extended N-terminus and a bent inwards C-terminus (Figure 5.6B,C,D,F) gave fits similar to those with a bent inwards N-terminus and an extended C-terminus (Figure 5.6A,E,G,H). The theoretical distance distribution curves $P(r)$ also showed good agreements with those of the experimental curves obtained previously (Osborne et al., 2018b).

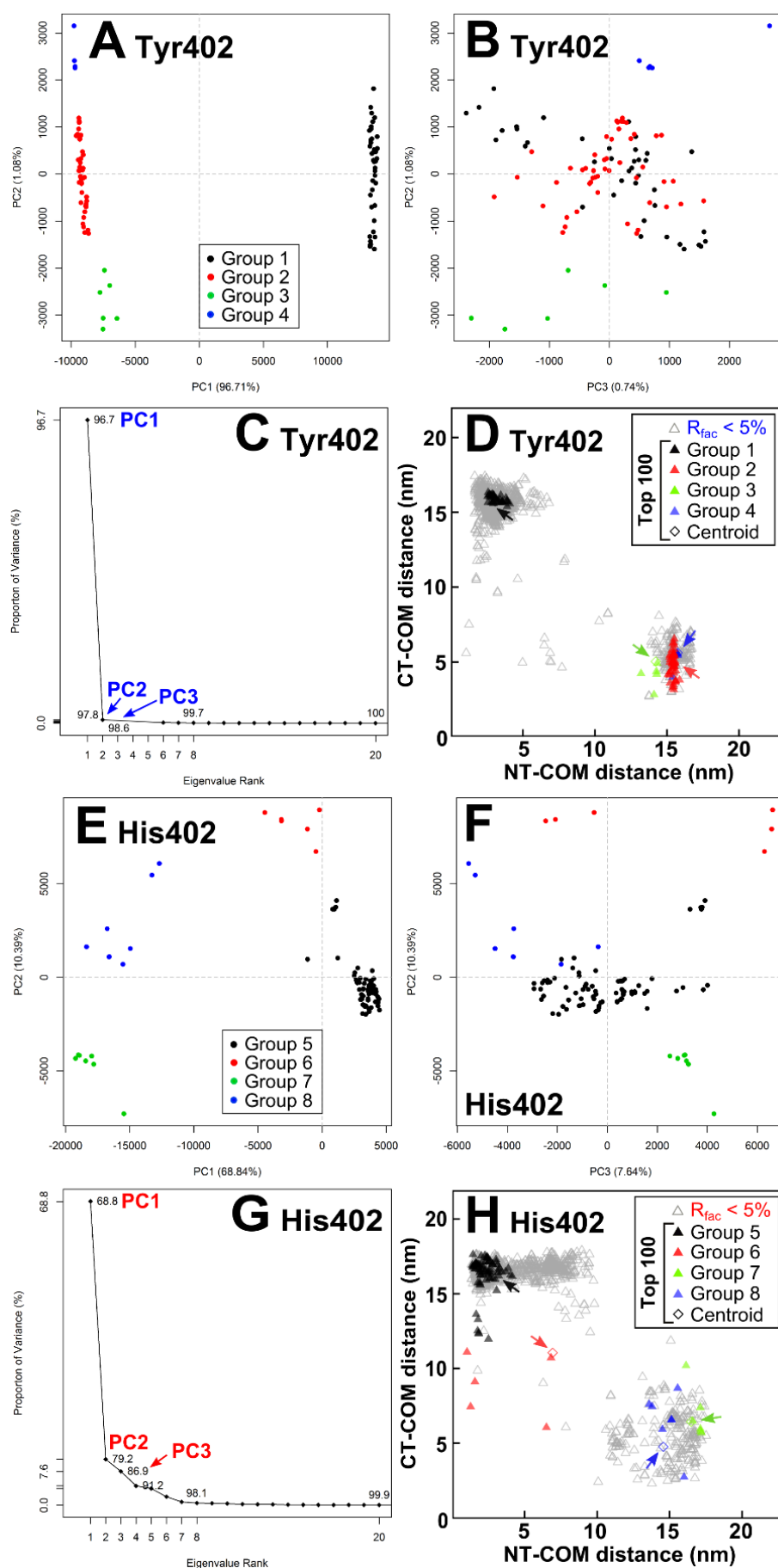


Figure 5.5 Principal component analyses of the 100 best-fit models for FH Tyr402 and FH His402. Figure legend overleaf.

Figure 5.5 (continued) Principal component analyses of the 100 best-fit models for FH Tyr402 and FH His402. For each of *A, B*, FH Tyr402 and *E, F*, FH His402, the 100 best-fit models were grouped by principal component analysis (PCA) into four groups 1-4 (black, red, green and blue, respectively) as exemplified by the first three principal components (PC2 vs. PC1 and PC3 vs. PC2). *C, G*, The first three eigenvalue rankings (PC1 to PC3) accounted for variances of 98.5% and 86.9% in the 100 best-fit FH models. *D, H*, The black, red, green and blue triangles correspond to the PCA groups 1-4 and 5-8 in each set of best-fit 100 models. The four arrowed diamonds corresponds to the average (centroid) model for each PCA group.

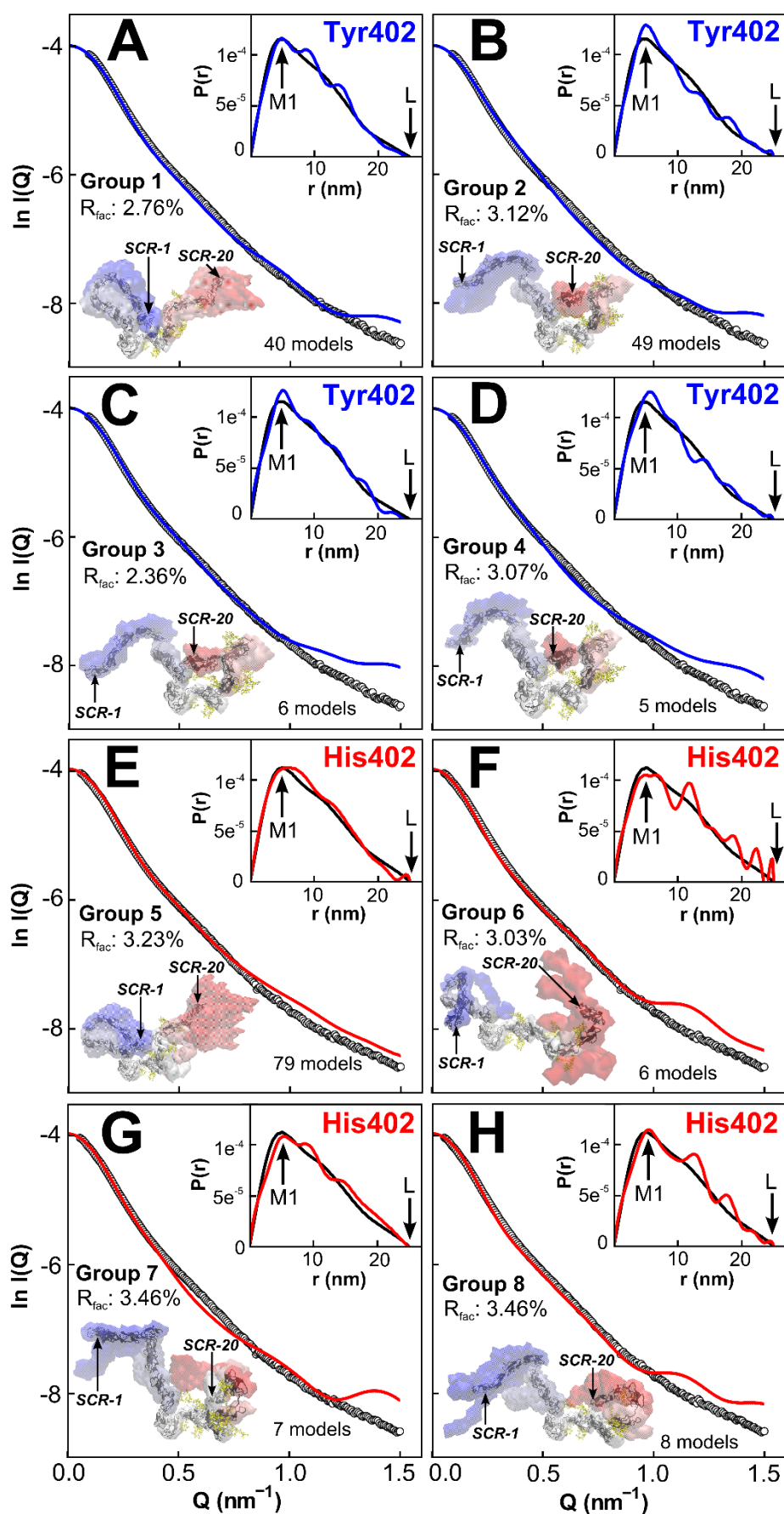


Figure 5.6 Scattering curve fits for the centroid PCA models for FH. Figure legend overleaf.

Figure 5.6 (continued) Scattering curve fits for the centroid PCA models for FH. For each of *A-D*, FH Tyr402 and *E-H*, FH His402, the four panels display the centroid FH model for the PCA groups 1-4 and 5-8 respectively in density plots. The number of FH models in each PCA group is displayed. The experimental curve is denoted by black circles and the theoretical curve is denoted as solid blue/red lines. The inset shows the $P(r)$ curves for the experimental data (black line) and modelled curve (blue/red line). For each PCA group, the conformational space is represented as blue (NT), white or red (CT) density. The average (centroid) model of each PCA group is depicted as a black cartoon in which SCR-1 and SCR-20 are arrowed.

Thus, the eight centroid FH models gave very similar M1 peak values that agreed with the experimental $P(r)$ curves obtained previously where M1 was 4.7-5.0 nm (Osborne et al., 2018b). Their intensities were comparable although these showed oscillations around the experimental $P(r)$ curves. Of note, the experimental $P(r)$ curves generally appeared smoother than the theoretical curves. This is likely due to both the lower resolution and the fluctuation between the alternative conformations of the solution structure studied by the experimental curves. Only eight of the 200 best-fit structures had conformations in which both the N-terminal and C-terminal regions were inwardly bent, and none showed that both regions were extended in any structure. The same PCA analyses for the other pair of available Tyr402 and His402 curves gave similar structural outcomes (not shown). The normalised Kratky analyses for the modelled best-fit curves (Figure 5.7) showed that, as for the experimental analyses in prior work (Osborne et al., 2018b), a clear peak was obtained that tailed off at large Q/R_G values, thus validating that FH is a multi-domain protein whose domains were connected by linkers (Receveur-Brechot & Durand, 2012).

5.5 Discussion

FH is an essential regulatory glycoprotein that protects host cells from complement activation both at the cell surface and in the fluid phase. In this present study, the most accurate molecular models to date have been produced for full-length FH, and these suggest that FH exists in two distinct conformations that enable its complement regulatory functions. Common and rare genetic variants in FH are associated with major and rare inflammatory diseases including AMD, Alzheimer's, aHUS and C3G (Osborne et al., 2018a). For AMD, each of the Tyr402His and Val62Ile variants are associated with increased and decreased risk, respectively. In order to perform its regulatory activities, FH binds to multiple physiological ligands, including to itself, with weak K_D values in the μM range, and has a highly glycosylated 20-domain structure (Figure 5.1). Up to now, the only FH molecular structures had been determined by preliminary modelling of X-ray scattering curves of heterozygous FH, resulting in two different folded-back structures whose significance was unclear (Nan et al., 2010; Okemefuna et al., 2009; Rodriguez et al., 2014). Here, by utilising a much-improved atomistic modelling procedure based on molecular dynamics and Monte Carlo procedures and statistical assessments of the outcome, similar detailed molecular models for both the Tyr402 and His402 allotypes of homozygous full-length FH, including its eight FH glycans, were determined. The statistical validation of two alternative FH conformations significantly

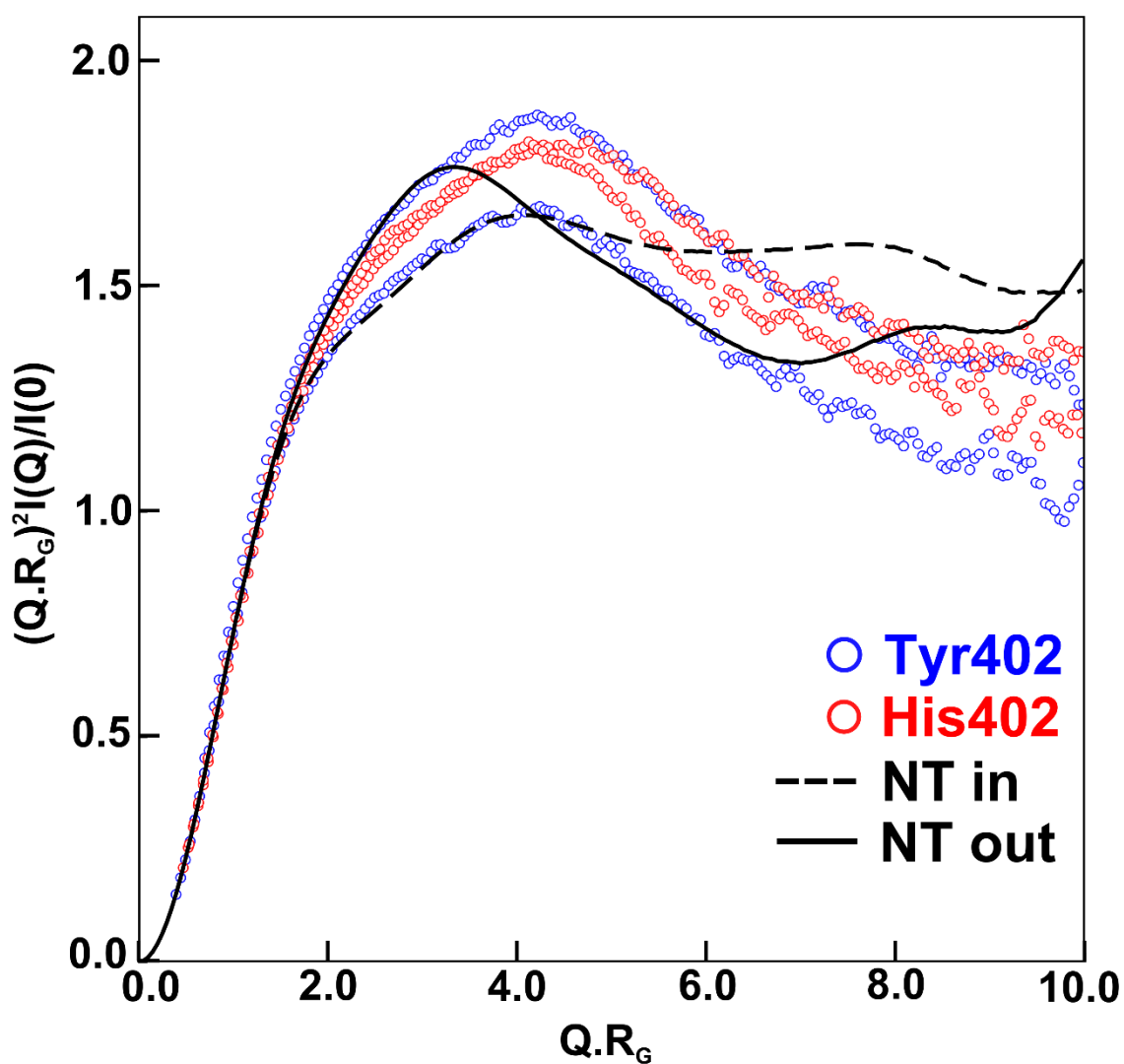


Figure 5.7 Normalised Kratky plot for the experimental and best-fit FH curves. For both allotypes, the theoretical curves for the most representative N-terminus extended and C-terminus extended FH models are shown as black solid and dashed lines, respectively. These correspond to the centroid models shown in Figures 5.6B and 5.6E respectively. The experimental curves are shown as blue and red open circles for the Tyr402 and His402 allotypes, respectively.

improves the understanding of ligand binding to FH and how this may be perturbed in its function.

5.5.1 Self-association of FH

Characterisation of FH self-association is important for both determining its conformational solution states and investigating the development of drusen in AMD. FH self-association is important for FH function because this will increase the local concentration of its ligand-binding sites. Biologically, given that the plasma concentration of FH is around 0.7 mg/ml (Perkins et al., 2010a), reversible dimers and trimers are expected to form *in vivo* (Osborne et al., 2018b). The resulting clustering of FH on cell surfaces, when bound to surface markers and C3b/C3d, is likely to enable broader surface protection, thus being functionally significant (Ferreira et al., 2010). However, the accumulation of FH by clustering may also relate to the formation of soft drusen in AMD via chronic inflammation, in particular if FH oligomer formation becomes irreversible (Hageman et al., 2001). Consistent with this, studies with full-length FH showed that its affinity for both heparin and HS were not altered by the Tyr402His polymorphism (Kelly et al., 2010; Ormsby et al., 2008; Toomey et al., 2018). It was also found that FH multimer formation is enhanced in the presence of zinc, which inhibits FH function (Nan et al., 2008b). In work conducted prior to this PhD thesis, it was shown using four independent methods that wild-type FH Tyr402 and AMD risk-associated FH His402 both significantly self-associate (Osborne et al., 2018b). These FH preparations were also genotyped for the AMD-protective Val62Ile polymorphism, which has a subtly better capacity to bind C3b, inhibit proconvertase formation and catalyse inactivation of C3b, by a minor structural rearrangement within SCR-1 (Hocking et al., 2008; Tortajada et al., 2009). SAXS analyses of the Guinier $I(0)/c$, R_G and R_{XS-1} parameters also showed clear concentration effects attributable to self-association (Osborne et al., 2018b). In order to model the monomeric FH structure for this thesis chapter, one immediate consequence of this was the need to allow for self-association effects in the SAXS structural studies. Thus, it was necessary to extrapolate the scattering curves to zero concentration.

5.5.2 Atomistic modelling of two FH conformations

The new molecular modelling of the FH scattering curves has benefitted from five improvements (Perkins et al., 2016): (i) the application of molecular dynamics to generate

an initial full FH structure; (ii) the use of 17 high resolution SCR structures, in distinction to up to 11-14 SCR structures previously; (iii) the inclusion of eight energy-refined glycan chains in FH; (iv) the use of rapid Monte Carlo simulations with the inter-SCR linkers to generate a large number of 29,715 physically realistic trial FH structures; (v) statistical analyses of the resulting ensemble of best-fit structures. Interestingly, two previous modelling studies in 2008 yielded alternative folded-back FH structures with either the N-terminal or C-terminal SCR domains in an extended conformation (Nan et al., 2010; Okemefuna et al., 2009; Rodriguez et al., 2014), yet their significance was unclear. Here, the statistical analysis of the 29,715 curve fits showed that both alternative FH conformations were valid outcomes from the SAXS modelling (Figure 5.8A,B). In the 200 best-fit structures, some 81 and 119 structures respectively corresponded to (i) an extended N-terminus and a bent inwards C-terminus, and (ii) a bent inwards N-terminus and an extended C-terminus. In principal, either of these two FH structures may be the actual solution structure, or both FH structures may co-exist as an equilibrium in solution. Consideration of C3b and C3d binding (below) suggested both FH conformations exist. Further statistical analyses of the 200 best-fit structures showed that the two FH Tyr402 and His402 allotypes displayed the same solution structures.

5.5.3 Biological significance of the FH Tyr402 and His402 models

To summarise FH activity, SCR-1/5 is involved in both cofactor and decay-acceleration activities (Kuhn & Zipfel, 1996; Sharma & Pangburn, 1996) whereas SCR-19/20 is involved in the recognition of host cell surfaces (Pangburn, 2000). These two functions are mediated by N-terminal SCR-1/4 binding to both C3b and FI, and C-terminal SCR-19/20 binding to C3d, heparin and sialic acid. C3dg and C3d, being two fragments of C3 that each contain the thioester domain (TED) for host-surface binding, are formed by the breakdown of C3b by FI and/or serum proteases and both bind to complement receptor 2 in order to trigger immune complex clearance and stimulate the immune response (Law & Reid, 1995). C3d is also an autologous helper T-cell target (Knopf et al., 2008). During complement regulation at host cell surfaces, an increase in the number of C3d binding sites for SCR-19/20 would enable more FH to be recruited at the host cell surface for a more rapid breakdown of C3b (Kajander et al., 2011). For cell surface attachment of FH, an optimal combination of affinities for each of C3b and polyanion ligand is essential for C-terminal FH binding (Ferreira et al., 2010), for which

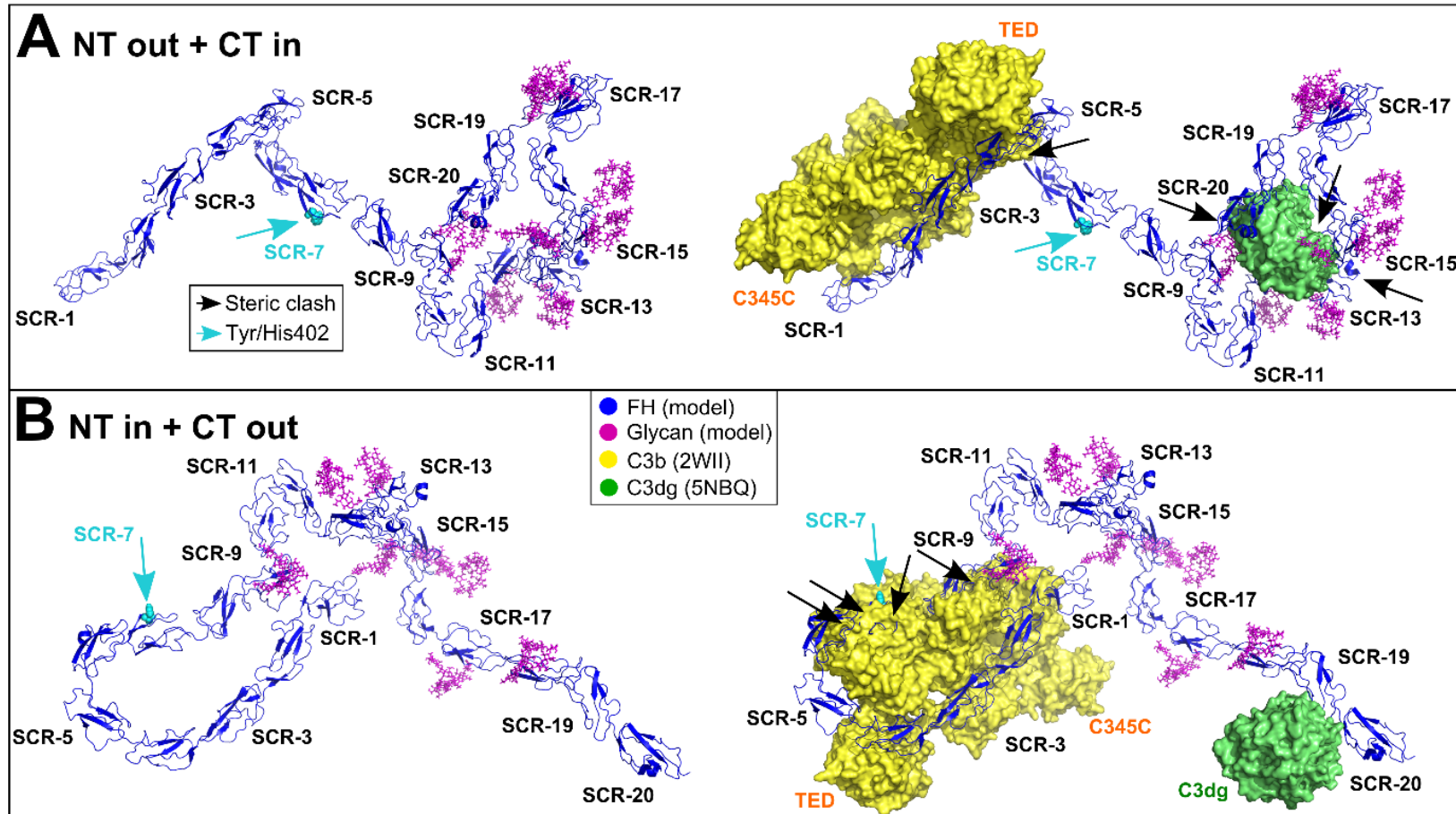


Figure 5.8 Best-fit FH centroid models superimposed onto C3b and C3dg. For both allotypes, the most representative A, N-terminus extended and B, C-terminal extended FH models are shown as blue ribbon traces (left). These correspond to the centroid models shown in Figures 5.12B and 5.12E respectively. The coordinates are available for download in Supplementary Materials. These models were superimposed onto the SCR-1/4-C3b and SCR-19/20-C3d crystal structures, in which C3b is shown as a yellow surface and C3dg is shown as a green surface (right; PDB codes: 2WII, 5NBQ). The eight glycans are shown in magenta. The cyan arrow indicates Tyr402/His402. Black arrows indicate steric clashes of FH with C3b or C3dg.

both SCR-7 and SCR-19/20 bind bivalently via heparin (Khan et al., 2012; Perkins et al., 2012).

The functions of FH were clarified by superimposing the two alternative best-fit conformations of FH with crystal structures for the SCR-1/4-C3b complex and the SCR-19/20-C3dg complex (PDB codes: 2WII, 5NBQ) (Kolodziejczyk et al., 2017; Wu et al., 2009).

(i) The N-terminal extended FH conformation accounted for binding to C3b at SCR-1/4, and to heparin and sialic acid, but less so for C3dg. The binding of C3b to SCR-1/4 showed minimal steric clashes with the rest of FH (black arrow, Figure 5.8A). Likewise, for binding to fluid phase FI, the SCR-1/3 domains were accessible, including the FH residues 101, 106, 112 and 119 that were important for this interaction (Wu et al., 2009). For bivalent SCR-7 and SCR-19/20 binding to heparin and/or sialic acid on cell surfaces, the essential residues in each of SCR-7 (Tyr/His402) and SCR-19/20 (FH residues 1181-1183, 1191, 1195, 1196, 1198 and 1199) were solvent-accessible. However, C3dg binding to SCR-19 resulted in significant steric clashes with SCR-9, SCR-12/15 and the glycans (black arrows, Figure 5.8A), implying that some structural rearrangement will be required for C3dg binding. During complement regulation at host cell surfaces, an increase in the number of C3d binding sites for SCR-19/20 would enable more FH to be recruited at the host cell surface for a more rapid breakdown of C3b (Kajander et al., 2011). This N-terminal extended FH conformation did not favour cell surface binding via C3d (and heparin), which may restrict it to the fluid phase for cofactor and DAA.

(ii) In contrast, the C-terminal extended FH model accounted for binding to C3dg at SCR-19 or SCR-20, and to heparin and/or sialic acid on cell surfaces, but not to C3b at SCR-1/4. Thus, the binding of C3b to SCR-1/4 resulted in large steric clashes with the SCR-6/8 domains (black arrows, Figure 5.8B). For binding to fluid phase FI, the SCR-1/3 domains were accessible. In order to bind C3 via SCR-1/4, potential steric hindrance from SCR-6/8 would need to be removed, most likely by moving the flexible linkers between SCR-4/5 and SCR-5/6 (Wu et al., 2009). Such large conformational changes in other complement proteins have already been reported for the TED domain in C3b, and the Bb fragment from C3b within the C3 convertase (Wu et al., 2009). The accessibility of the Tyr/His402 residue in SCR-7 was not compromised. For bivalent binding to heparin and/or sialic acid on cell surfaces, each of SCR-7 and the essential residues in SCR-19/20 (above) were accessible. For C3dg binding to SCR-19 or SCR-20 (Morgan et

al., 2011), no steric clash was seen (Figure 5.8B). The shortest distance between the SCR-1/4 and SCR-6/9 domains was ~6 nm between Arg78 in SCR-1 and Asp538 in SCR-9. These residues bear opposite charges (Armstrong et al., 2016). I speculate that these domains may electrostatically attract each other; indeed conformational changes in FH with increase in ionic strength have been reported (Okemefuna et al., 2009). Further studies are required to investigate such a proposed mechanism.

The existence of two FH conformations with different C3b and C3d binding strengths accounts for previously-reported K_D values for each of SCR-1/4 binding to C3b (10-14 μ M) and SCR-19/20 binding to C3b or C3d (1-5 μ M and 0.2-8 μ M respectively). These K_D values are similar to the K_D value of full-length FH binding to C3b (0.6-3 μ M) (Perkins et al., 2012), indicating no synergy between the two binding sites at SCR-1/4 and SCR-19/20, and meaning that these two C3b and C3d sites are independent of each other. This outcome is in concord with N-terminal extended and C-terminal extended FH models which bind to C3b or C3d separately, but not both together. In summary, FH is seen with an N-terminal extended conformation that is able to both dismantle fluid phase C3 convertases and perform cofactor activity, and also with a distinct C-terminal extended conformation that has increased affinity for binding to cell surfaces via both heparin and deposited C3d for enhanced local protection from complement activation. This is because the presence of C3d on the cell surface may increase the affinity of FH in the correct conformation for cell surface binding (Kajander et al., 2011). For the C-terminal extended FH structure, following its binding to cell surfaces, a conformational change involving the flexible linkers between SCR-4/5 and SCR-5/6 would allow SCR-1/4 to bind C3b. By this, C3b can either be inactivated to iC3b at host cell surfaces (cofactor activity) or separated from Bb in the C3 convertase (DAA). Overall, both alternative conformations are able to bind to glycosaminoglycans at the host cell surface, thus protecting this against excess C3b activity. However, the second structure with an extended C-terminus is more likely to be recruited to cell surfaces by the proposed increased affinity for C3d. In summary, the SAXS data modelling has resulted in the proposal of two distinct FH conformations that would enable FH to perform separate C3b or C3d binding functions in either the fluid-phase or at host cell surfaces respectively.

Chapter Six

**Structural analyses rationalise the
distribution of aHUS and C3G rare
variants in complement factor H**

6.1 Summary

Atypical haemolytic uraemic syndrome and C3 glomerulopathy (aHUS, C3G) are two severe rare diseases associated with complement dysregulation. The most common variants were missense changes in the regulators factor H, factor I and membrane cofactor protein (FH, FI and MCP), and the activators C3 and factor B (FB). Their molecular correlations with aHUS and C3G were assessed by mapping the variants onto three-dimensional structures for the FH-C3b-FI, FH-C3dg and C3b-FB complexes and full-length FH. The 128 aHUS rare missense variants in FH mostly affected three types of residues, namely 13 case alleles in 42 buried residues (31%), followed by 290 case alleles in 1045 non-binding surface-accessible residues (28%), and finally 12 case alleles in 92 residues at C3b-binding interfaces (13%), but none were found at the FI interfaces. In contrast, for 19 C3G variants in FH, these three types of residues totalled 2% (one case allele), 3% (26 case alleles) and 5% (five case alleles) respectively, and two case alleles (6%) within the FI binding interface were now involved. Another prominent group of variants involve Cys residues involved in disulphide bridges. Only those in the regulators FH, MCP and FI were associated with aHUS, while only those in FH and C3 were associated with C3G, when these were compared to the ExAC reference dataset. In conclusion, these differences between aHUS and C3G reflect their different pathologies, and importantly indicate that the structural location of a newly-discovered variant may predict the occurrence of aHUS or C3G in patients.

6.2 Introduction

In the complement AP, aHUS is associated with defects in cell surface regulation that lead to TMA, whereas C3G is associated with fluid phase dysregulation which leads to abnormal complement deposition in glomeruli (Goodship et al., 2017). aHUS was associated with rare genetic variation in FH, MCP, FI, C3 and FB, while C3G was associated with rare variation in FH and C3 only (Osborne et al., 2018a). For both aHUS and C3G, predisposing RVs lead to either a LoF in the regulators (FH, MCP or FI), or a GoF in the activator proteins (C3 or FB), and AP dysregulation. Protein three-dimensional structures are known for many of the AP proteins. This includes crystal structures for FH SCR-1/4-C3b (Wu et al., 2009), FH SCR-1/4-FI-C3b (Xue et al., 2017) and FH SCR-19/20-C3dg (Kolodziejczyk et al., 2017), a solution structure for glycosylated full length FH (Osborne et al., 2018b), and a crystal structure for C3b-FB (Forneris et al., 2010). In terms of general protein function and structure, residues involved in molecular interactions are surface-associated whereas residues involved in structural stabilisation are buried or surface-associated. In order to predict these structural residue types, residue positions can be determined by submitting protein structural co-ordinates to surface accessibility or “accessible surface area” calculations (Lee & Richards, 1971). In order to predict which surface-accessible residues are involved in binding interfaces, such as for the regulatory C3b-FH-FI or activating C3b-FB complement complexes, the buried surface area can also be calculated.

The SCR is the most abundant domain in the complement proteins and occurs in FH, MCP and FB. SCR domains generally contain about 60 residues arranged as six to eight β -strands numbered β 1 to β 8 with a hypervariable sequence loop between strands β 2 and β 3. SCR domains are structurally stabilised by two disulphide bridges formed by two pairs of conserved Cys residues (Chapter 3, Section 3.3.8). SCR domains also contain one conserved Trp residue. In all proteins, disulphide bonds are the second most common covalent link between amino acids and are estimated to be present in ~10% of mammalian proteins. Disulphide bonds form in the oxidising environment of the endoplasmic reticulum. In addition to structural stability, disulphide bonds can mediate thiol-disulphate interchange reactions in substrate proteins (catalytic) and control protein function by triggering a conformational change upon being broken or formed (allosteric) (Schmidt et al., 2006). For any protein, the disruption of a structural disulphide bond often leads to LoF via protein misfolding. In genetic disease, these LoF events can be predicted

by the presence of predisposing rare missense variants in disulphide bond-forming Cys residues. For FB, an additional five disulphide bonds are present in the aL, von Willebrand Factor A and serine protease domains (Forneris et al., 2010). For C3 and FI, there are ten and 20 structural disulphide bonds present, respectively (Xue et al., 2017). Previously for FH and MCP, SCR consensus domain analyses associated missense variants in disulphide bridge-forming Cys residues with aHUS and AMD (Rodriguez et al., 2014). For example, the two Cys-affecting FH variants (p.Cys518Arg and p.Cys914Tyr) were identified in a FH-deficient patient with a C3G-like phenotype for which FH was synthesised but retained in the endoplasmic reticulum (Ault et al., 1997; Schmidt et al., 1999).

The aim of this thesis chapter was to bring together the aHUS and C3G RV analyses from Chapter 4 and the FH modelling from Chapter 5 in order to identify whether the structural location of a variant in FH complexes may predict the occurrence of aHUS or C3G in patients. Here, the structural basis for the effect of RVs that predispose for aHUS and C3G were reviewed by predicting their functional impacts such as binding sites or protein stability. Firstly, the RV distributions in the complexes of both FH SCR-1/4-C3b (Wu et al., 2009) and FH SCR-19/20-C3dg (Kolodziejczyk et al., 2017), with improved structural models for glycosylated full length FH (Chapter 5) (Osborne et al., 2018b), and crystal structures for FH SCR-1/4-FI-C3b (Xue et al., 2017), and C3b-FB (Forneris et al., 2010) were analysed. Secondly, in order to analyse the abundances of the aHUS and C3G RVs in the functional regions of complement SCR domains, an SCR consensus domain was defined based on averaging the known structures of SCR domains (Rodriguez et al., 2014). This consensus was here updated for FH and MCP. For both aHUS and C3G, the frequencies of RVs in the so-called hypervariable loop were not significantly greater than for the rest of the SCR domain. For FH, the aHUS variants mostly affected FH residues which were surface-inaccessible (or buried), followed by non-binding surface-accessible residues and finally those at the C3b-binding interfaces. In contrast, for C3G, in addition to these three types all at similar frequencies to each other, those within the FI binding interface were now involved. Thirdly, for the SCR domains in FH and MCP, the frequency of missense rare variation in the four conserved Cys residues that mediate disulphide bridges was significantly greater for aHUS than for the ExAC reference genomic dataset (Lek et al., 2016). This result was also observed for aHUS for Cys residues involved in disulphide bridges in FI, but not for C3 or FB. For C3G, this was only observed for FH and interestingly, for C3. Overall, disulphide bridge

disruption was associated with aHUS in the AP regulators FH, MCP and FI, and, in contrast, with C3G for FH and C3. In summary, these three considerations suggest that the structural locations of complement variants may predict the occurrence of aHUS or C3G in patients.

6.3 Methods

For the aHUS, C3G and ExAC datasets, rare missense variants in the five genes *CFH*, *C3*, *CFI*, *CD46* and *CFB* were filtered and extracted from the Database of Complement Gene Variants by using SQL commands as previous (Osborne et al., 2018a).

6.3.1 Rare variant distributions in FH, C3, FI and FB complexes

For the filtering, the rare missense variants for the aHUS and C3G datasets were only included if they were classified as pathogenic, likely pathogenic or uncertain significance (Chapter 4, Section 4.3.6) and were not located in the signal peptide. The total numbers of variants correspond to those previously analysed (Osborne et al., 2018a).

The N-terminal extended and C-terminal extended FH structures were aligned with C3b-FH SCR-1/4 (PDB code: 2WII) (Wu et al., 2009) and FI-C3b-SCR-1/4 (PDB code: 5O32) (Xue et al., 2017), and C3dg-FH SCR-19/20 (PDB code: 5NBQ) (Kolodziejczyk et al., 2017), respectively, by using PyMol commands (Chapter 5, Section 5.5.2). For the C3b and FB convertase complex, the crystal structure (PDB code: 2XWJ) (Forneris et al., 2010), was used. All the sequences were converted to HGVS numbering using PyMol commands. The rare missense variants extracted for aHUS, C3G and both aHUS and C3G were then mapped to their respective α -carbon atoms as red, yellow and black spheres, respectively, onto the structures for FH, C3, FI and FB by using PyMol commands.

6.3.2 Consensus SCR domain analyses

The consensus sequence from 124 human complement SCR sequences was computed as previous (Rodriguez et al., 2014) (Figure 6.1). This included the averaged side-chain accessibilities of the 27 experimental SCR structures aligned to the consensus

structure which were computed from DSSP software for secondary structure assignments and side-chain surface accessibilities (Kabsch & Sander, 1983).

For aHUS and C3G, rare missense variants were only included if they were classified as pathogenic, likely pathogenic or uncertain significance (Chapter 4, Section 4.3.6) and were not located in the signal peptide. The total numbers of variants corresponded to those previously analysed (Osborne et al., 2018a).

The frequency of affected (and unaffected) residues was calculated by dividing the numbers of affected (and unaffected) residues by all residues (Table 6.1). In order to compare the frequencies of the aHUS and C3G rare missense variants between regions, the Fisher's two-tailed exact test for a 2×2 contingency table was used. The null hypothesis stated that there was no difference in frequency and the alternative stated that there was a difference in frequency. The significance level was 0.05. GraphPad QuickCalcs (<https://www.graphpad.com/quickcalcs/>) was used.

In order to consider how many times each variant had been seen in aHUS and how many patients were screened in total, these analyses were repeated with AF data (Table 6.2). By this, both the total allele count and the total AF of aHUS rare missense variants was calculated for each region of interest (Table 6.2). The aHUS AF takes into account the total number of alleles screened (i.e. two per patient) for each gene, this being 6256 *CFH* alleles and 5884 *CD46* (MCP) alleles. In order to normalise these total allele counts and AFs by the size of the region (i.e. size), they were divided by the number of residues in the region. These analyses were not performed for C3G because there were too few variants thus the frequencies were too low to give reliable results.

6.3.3 Surface-accessibility analyses of FH residues

For full-length FH in complex with FI and C3b, in order to determine whether the FH residues become buried, remain surface-exposed or are involved in binding interfaces, the residue surface accessibilities were calculated. This was done by using the crystal structures of C3b-FH SCR-1/4 (PDB code: 2WII) and FI-C3b-SCR-1/4 (PDB code: 5O32) (Figure 6.1B) and the software Protein Interfaces, Surfaces and Assemblies (PISA) at the PDBePISA server (<http://www.ebi.ac.uk/pdbe/pisa/>) (Krissinel & Henrick, 2007). By this, only the interactions and macromolecular contacts already present in the crystal

structures were analysed. These contacts were superimposed with full-length FH as part of FH SCR-1/4. PISA calculates the accessible surface area by rolling a water probe of 1.4Å in diameter over the surface of the protein and summing up all sampled points in contact with this probe, which then represents the surface area, normalised by the precision of the element analysis. For two residues, the interface between them or buried surface area is calculated from their accessible surface areas. For each FH residue, the aHUS and C3G rare missense variant allele counts and frequencies were calculated by using the Database of Complement Gene Variants with SQL commands, as previous (Osborne et al., 2018a) (Table 6.3).

6.3.4 Frequencies of rare variants in structural disulphide bridges

For each of the five proteins (FH, FI, MCP, FB and C3), rare missense variants for the aHUS, C3G and ExAC buried Cys datasets were only included if they affected disulphide bond forming-Cys residues and had an ExAC AF < 0.01%. By using the ExAC AF filter for all of the aHUS, C3G and ExAC datasets, a potential bias was introduced whereby common variants in aHUS and C3G were included but common variants in ExAC were excluded. However, for these analyses, there were no common variants affecting Cys residues for ExAC (all Cys variants were rare), and the variants affecting Cys residues that were the most common in the aHUS and C3G datasets were also rare. Overall, it was concluded that the effects of this potential bias were minimal.

For the SCR domains in FH, MCP and FB, all their Cys residues were involved in disulphide bridges. Thus, SCR domains account for all of FH, 70% of MCP (SCR-1/4; 251/358 residues) and 26% of the activator FB (SCR-1/3; 195/764 residues). For the remaining non-SCR domain residues of C3, FB and FI, the number of disulphide bridge-forming Cys residues was extracted from their structural co-ordinate files (PDB codes: 2XWJ and 5O32). Thus, for FB, an additional five disulphide bonds were present in the aL, von Willebrand Factor A and serine protease domains (Forneris et al., 2010). For C3 and FI, ten and 20 structural disulphide bonds were present, respectively (Xue et al., 2017).

For each protein, for each of the aHUS and C3G datasets, the total AF of rare missense variants that affected disulphide bridge-forming Cys residues was calculated (Table 6.4). This was calculated by dividing the total allele count for the variants by the

total number of alleles screened for the gene. For ExAC, this was calculated by summing the allele count and dividing this by the mean allele number, as previous ([Chapter 4](#)).

In order to compare the frequencies of rare missense variants that affected disulphide bridge-forming Cys residues in aHUS and C3G with those for the reference ExAC dataset, the Chi-square test (χ^2) with Yates' correction for a 2×2 contingency table was used. The null hypothesis stated that there was no difference in AF. The alternative stated that there was a difference in frequency. The significance level was 0.05. GraphPad QuickCalcs (<https://www.graphpad.com/quickcalcs/>) was used. Prior to each test, the power was calculated by using the program PS Power and Sample Size Calculations (Version 3.0) ([Dupont & Plummer, 1990](#)). The tests for FH and MCP showed powers of at least 80%.

6.4 Results

6.4.1 Frequency of aHUS and C3G rare variants in hypervariable loop

In order to appreciate the impact of the rare missense variants on the SCR domain, the updated SCR consensus domain for FH and MCP is shown in [Figure 6.1](#) for aHUS and C3G. The averaged side-chain accessibility of the 27 experimental SCR structures aligned to the SCR consensus structure is also shown ([Rodriguez et al., 2014](#)). In this present study, a total of 170 rare missense variants for aHUS and C3G were mapped to their positions within the consensus sequence. These variants corresponded to totals of 115 FH and 37 MCP for aHUS (red and blue texts, respectively), ten FH for C3G (yellow text), and eight FH that were identified in both aHUS and C3G patients (black text) ([Figure 6.3](#)).

For aHUS, while the rare missense variants were randomly distributed throughout the consensus SCR domain, the five conserved Cys and Trp residues were the most populated with 40 occurrences out of 160 (25%). The loss of any of these residues appears to impact the stability of the SCR domain. For C3G, the distribution of the rare missense variants also appeared to be random. However, out of the four conserved Cys residues, only the third conserved Cys residue was affected by C3G variants (three; p.Cys431Tyr, p.Cys915Ser, p.Cys1218Arg) and all of these were also identified in aHUS patients ([black](#)

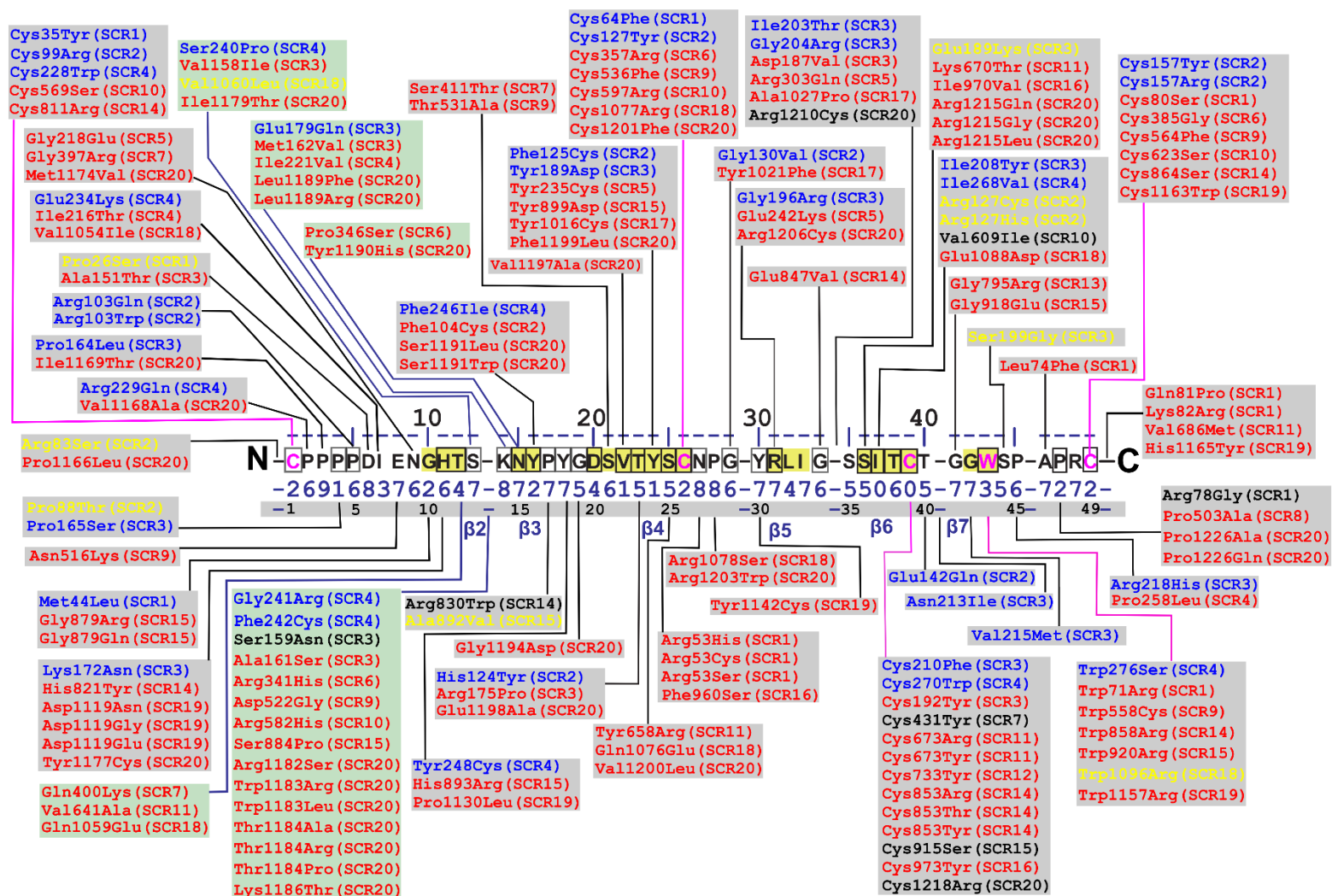


Figure 6.1 Location of aHUS and C3G rare missense variants on a SCR consensus sequence. Figure legend overleaf.

Figure 6.1 (continued) Location of aHUS and C3G rare missense variants on a SCR consensus sequence. The consensus SCR sequence for FH and MCP shows the most commonly occurring residue at each position. Mapped to their positions within the consensus sequence are the total of 170 rare missense variants for aHUS and C3G. These correspond to 115 FH and 37 MCP for aHUS (red and blue, respectively), 10 FH for C3G (yellow), and 8 FH that were identified in both aHUS and C3G patients (black). On the SCR consensus sequence, the residues corresponding to the six main β -strands $\beta 2$ to $\beta 7$ in SCR domains are shown in yellow. For the hypervariable loop between $\beta 2$ and $\beta 3$ which corresponds to five residues 12–16, the 27 aHUS, 1 C3G and 1 identified in both aHUS and C3G rare missense variants are displayed against a green background. The consensus solvent side-chain accessibilities computed by DSSP are numbered from 0 to 9, where 0 denotes 0–9% accessibility, 1 denotes 10–19% accessibility, etc. Residues with values of 0 and 1 are considered to be buried in the protein structure.

text, Figure 6.1). This suggested an overlap in aHUS and C3G phenotypes at this Cys residue. In addition for C3G, the conserved Trp residue was affected by one variant.

For aHUS, 28 of the 160 rare missense variants (for FH and MCP), which corresponded to 24 distinct residues, occurred in the five consensus residues 12-16 of the hypervariable loop between β -strands β 2 and β 3 (blue, red and black text in green boxes, Figure 6.3) These five consensus residues of the hypervariable loop corresponded to 192 FH and MCP distinct residues in SCR domains. For the hypervariable loop, the frequency of residues affected by aHUS rare missense variants was 13% (Table 6.1). When this was compared to the frequency of 9% for the non-hypervariable loop regions, there was no significant difference ($p=0.1494$). However, for the four conserved Cys residues (96 residues in total), which were included in the ‘non-hypervariable loop’ region, the frequency of aHUS-affected residues (29; 30%) was significantly greater than for the remaining ‘non-Cys’ residues (8%; $p<0.0001$). For the non-hypervariable loop residues, when these four conserved Cys residues were taken out, the frequency of aHUS-affected residues was also 8% (last row, Table 6.1). The frequency of hypervariable loop residues, 13%, was now significantly greater than the frequency for both the non-hypervariable loop and non-Cys residues (8%; $p=0.0230$), but not as significant as for the four conserved Cys residues. This statistical analysis shows that alterations in the four Cys residues were indeed a major factor leading to aHUS.

In order to consider how many times each variant was observed in aHUS and how many patients were screened in total for *CFH* and *CD46* variants, these analyses were repeated with aHUS AF data (Table 6.2). By this, both the total allele count and the total AF of aHUS rare missense variants were calculated for the hypervariable loop residues, the four conserved Cys residues, and those residues not in either the hypervariable loop nor the four Cys residues (Table 6.2). These were then normalised by the size of each region by dividing by the number of residues for that region. For aHUS, the total allele counts were 52, 62 and 284 for the hypervariable loop, the four conserved Cys residues and neither the hypervariable loop nor the Cys residues, respectively. These corresponded to total AFs of 0.84%, 1.00% and 4.61%, respectively. After normalisation, the AFs were 0.0044%, 0.0104% and 0.0039%, respectively. Overall, the four conserved Cys residues had the highest normalised allele count and AF of 65% and 0.0104%, respectively. The other residues had very similar aHUS AFs ranging from 0.0039% to 0.0044% (Table 6.2). In summary, as noted above, the four conserved Cys residues (0.0104%) were

Table 6.1 Frequencies of aHUS and C3G affected residues in the consensus SCR

Location of residues	Total residues ^a	Affected residues (frequency ^b)		Unaffected residues (frequency ^b)	
		aHUS	C3G	aHUS	C3G
Hypervariable loop	192	24 (13%)	2 (1%)	168 (87%)	189 (99%)
Non-hypervariable loop	1272	117 (9%)	17 (1%)	1155 (91%)	1249 (99%)
<i>Cys</i>	96	29 (30%)	3 (3%)	67 (70%)	93 (97%)
<i>Non Cys</i>	1368	112 (8%)	16 (1%)	1256 (92%)	1345 (99%)
All	1464	141 (10%)	19 (1%)	1323 (90%)	1438 (99%)
Non-hypervariable loop and non-Cys	1176	88 (8%)	14 (1%)	1088 (92%)	1156 (99%)

^a “Total residues” was calculated by adding up the number of residues for each of the four SCRs in MCP and 20 SCRs in FH, including the linker residues.

^b The frequency of affected (or unaffected) residues was calculated by dividing the numbers of affected (or unaffected) residues by all residues.

Table 6.2 Allele frequencies of aHUS affected residues in the consensus SCR

Location of residues	Total residues ^a	aHUS allele count	aHUS allele count by residues (%) ^b	aHUS allele frequency (%)	aHUS allele frequency by residues ^c (%)
Hypervariable loop	192	52	27	0.84	0.0044
<i>Cys</i>	96	62	65	1.00	0.0104
Non-hypervariable loop and non-Cys	1176	284	24	4.61	0.0039
All	1464	398	27	6.45	0.0044
<i>Non-hypervariable loop</i>	1272	346	27	5.61	0.0044

^a “Total residues” was calculated by adding up the number of residues for each of the four SCRs in MCP and 20 SCRs in FH, including the linker residues.

^b Divided by the total number of residues in each region (second column).

^c The aHUS allele frequency takes into account the total number of alleles screened (i.e. two per patient) for each gene, this being 6256 *CFH* alleles and 5884 *CD46* alleles.

significantly more affected by aHUS rare missense variants, but the hypervariable loop residues (0.0044%) were not. For C3G, there were too few C3G variants giving very low frequencies (1% to 3%), thus the power was too low for these analyses.

6.4.2 Distributions of aHUS and C3G rare variants in complement complexes

For each of the five proteins (FH, FI, MCP, FB, C3), datasets of aHUS and C3G rare missense variants that were classified as pathogenic, likely pathogenic or uncertain significance ([Chapter 4, Section 4.3.6](#)) and not located in the signal peptide were extracted from the Database of Complement Gene Variants. These variants were categorised as either aHUS, C3G or both aHUS and C3G, and were mapped onto the protein structures by using red, yellow and black spheres, respectively ([Figure 6.2](#)).

Only rare missense variants that were classified as pathogenic, likely pathogenic or uncertain significance and were not located in the signal peptide were included. Thus, for FH, there were 128 aHUS (109 unique residues) and 19 C3G (18 unique residues) rare missense variants, of which eight in each dataset were common to both aHUS and C3G (8 unique residues). These are taken from column 1 of [Figure 4.3B](#). For C3, there were 61 aHUS (56 unique residues) and 24 C3G (24 unique residues) rare missense variants, of which five in each dataset were common to both aHUS and C3G (five unique residues). These are taken from the fourth protein of [Figure 4.3B](#). These affected residues were mapped onto the N-terminus extended FH model ([Chapter 5, Section 5.5.2](#)) in complex with C3b ([Wu et al., 2009](#)) ([Figure 6.1A](#)) and the C-terminus extended FH model in complex with C3dg ([Kolodziejczyk et al., 2017](#)) ([Figure 6.1B](#)). As previously observed ([Chapter 4, Section 4.5.4](#)), the aHUS rare missense variants ([red and black spheres, Figures 6.2A, B and C](#)) were clustered at SCR-20 and also distributed throughout the other SCR domains. For SCR-20, these variants likely affect heparin binding required for host-cell surface protection. For SCR-7, the three rare missense variants may also affect heparin binding. For the N-terminal domains, despite there being fewer aHUS variants than in the C-terminal domains ([Chapter 4, Section 4.4.8](#)) the aHUS variants within SCR-1/4 appeared to cluster at the interfaces between FH SCR-3 and C3b MG2/CUB, and FH SCR-4 and C3b TED/MG1 ([red and black spheres, Figure 6.2A and B](#)). For C3b, within the FH SCR-3/C3b MG2 interface, the MG2 variant p.Arg161Trp ([red arrow, Figure 6.2A and B](#)) had the greatest aHUS AF (1.1%) in the aHUS dataset, being identified in 52 aHUS cases in heterozygous form. For the FH SCR-3/C3b CUB interface

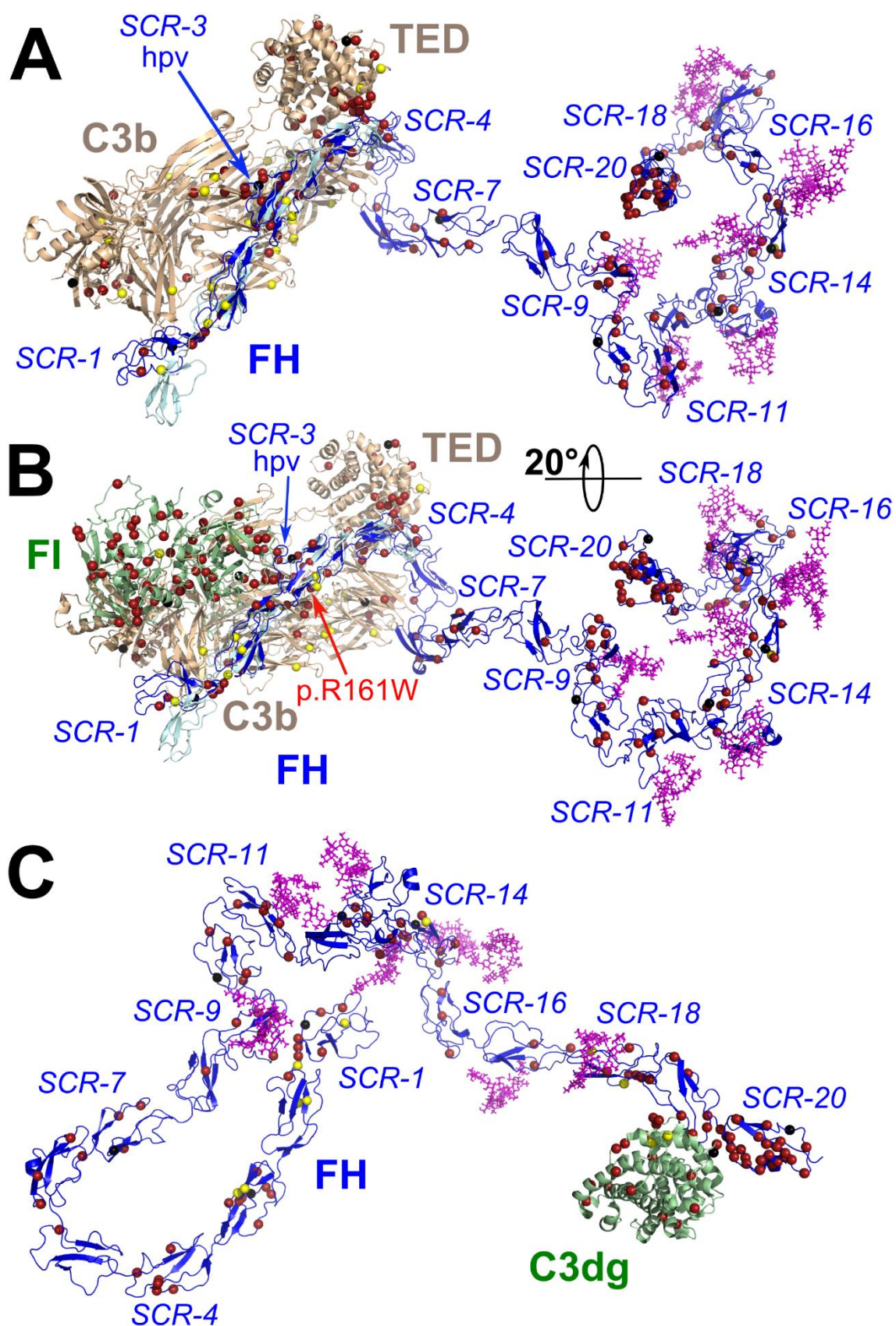


Figure 6.2 Rare missense variants mapped onto models for FH in complex with C3b and FI. Figure legend overleaf.

Figure 6.2 Rare missense variants mapped onto models for FH in complex with C3b and FI. Rare missense variants for aHUS, C3G and both aHUS and C3G mapped as red, yellow and black spheres, respectively, onto structural models shown in cartoon style by using PyMol. For FH, there were 128 aHUS (109 unique residues) and 19 C3G (18 unique residues) rare missense variants, of which eight in each dataset were common to both aHUS and C3G (eight unique residues). For C3, there were 61 aHUS (56 unique residues) and 24 C3G (24 unique residues) rare missense variants, of which five in each dataset were common to both aHUS and C3G (five unique residues). The N-terminal extended FH model was aligned with each of (A) C3b (tan) and FH SCR-1/4 (cyan) (PDB code: 2WII), and (B) FI (green), C3b (tan) and SCR-1/4 (cyan) (PDB code: 5O32). The variant p.Arg161Trp (red arrow) had the greatest aHUS allele frequency (1.1%) in the aHUS dataset. The hypervariable loop in FH SCR-3 is labelled as hpv (blue text). (C) The C-terminal extended FH structure was aligned with C3dg (green) and FH SCR-19/20 (not shown) (PDB code: 5NBQ).

(Wu et al., 2009), the FH hypervariable loop (hvp; residues 157 to 163) harboured the four aHUS missense variants p.Val158Ile, p.Ser159Asn, p.Alal61Ser, p.Met162Val, with aHUS AFs of 0.05%, 0.02%, 0.08% and 0.02% respectively. Interestingly, p.Ser159Asn was also seen in one C3G patient, suggesting an overlap in phenotypes at this position. The mutational frequency of the hypervariable loop in aHUS and C3G was further analysed using statistics in [Section 6.4.3](#).

For C3G, for FH and C3b, the rare missense variants were distributed mainly in C3 and in FH SCR-1/3 ([yellow and black spheres, Figure 6.2A, B and C](#)). This appeared to include the FH SCR-3 and C3b MG2/CUB binding interfaces. However, in contrast to aHUS, were no C3G rare missense variants in FH SCR-4 or the interfaces between FH SCR-4 and C3b TED or MG1. Also in contrast to aHUS, few C3G rare missense variants were seen in the middle and C-terminal SCR domains, as noted previously ([Osborne et al., 2018a](#)). For the SCR-7 and SCR-20 domains involved in binding heparin, the one C3G variant in SCR-7 and two C3G variants in SCR-20 were also seen in aHUS patients ([black spheres, Figure 6.2A, B and C](#)). For C3b, the two C3G variants p.Arg1303His and p.Arg1320Gln were located in the first and second cleavage sites thus were predicted to provide resistance against C3b cleavage by FI and a cofactor.

For FI, there were 65 aHUS (61 unique residues) and five C3G (five unique residues) rare missense variants, with four (four unique residues) common to both aHUS and C3G. The affected FI, FH and C3b residues were mapped onto the structure of FI complexed with C3b ([Xue et al., 2017](#)) and the N-terminus extended FH model ([Chapter 5, Section 5.5.2](#)) ([green, Figure 6.2B](#)). In contrast to FH and C3b, the aHUS affected residues for FI were randomly distributed over the whole structure, as observed previously ([Chapter 4, Section 4.4.10](#)).

For FB, there were 19 aHUS (19 unique residues) and seven (seven unique residues) C3G rare missense variants, with one variant common to both aHUS and C3G. The affected FB and C3 residues were mapped onto the structure of FB complexed with C3b ([Figure 6.3](#)). No clustering of variants was observed in the binding sites for the C3 convertase complex. However, interestingly, one C3G variant (p.Cys1518Arg) affected a disulphide bond-forming Cys residue in the C3 C345C domain ([cyan sphere, Figure 6.3](#)).

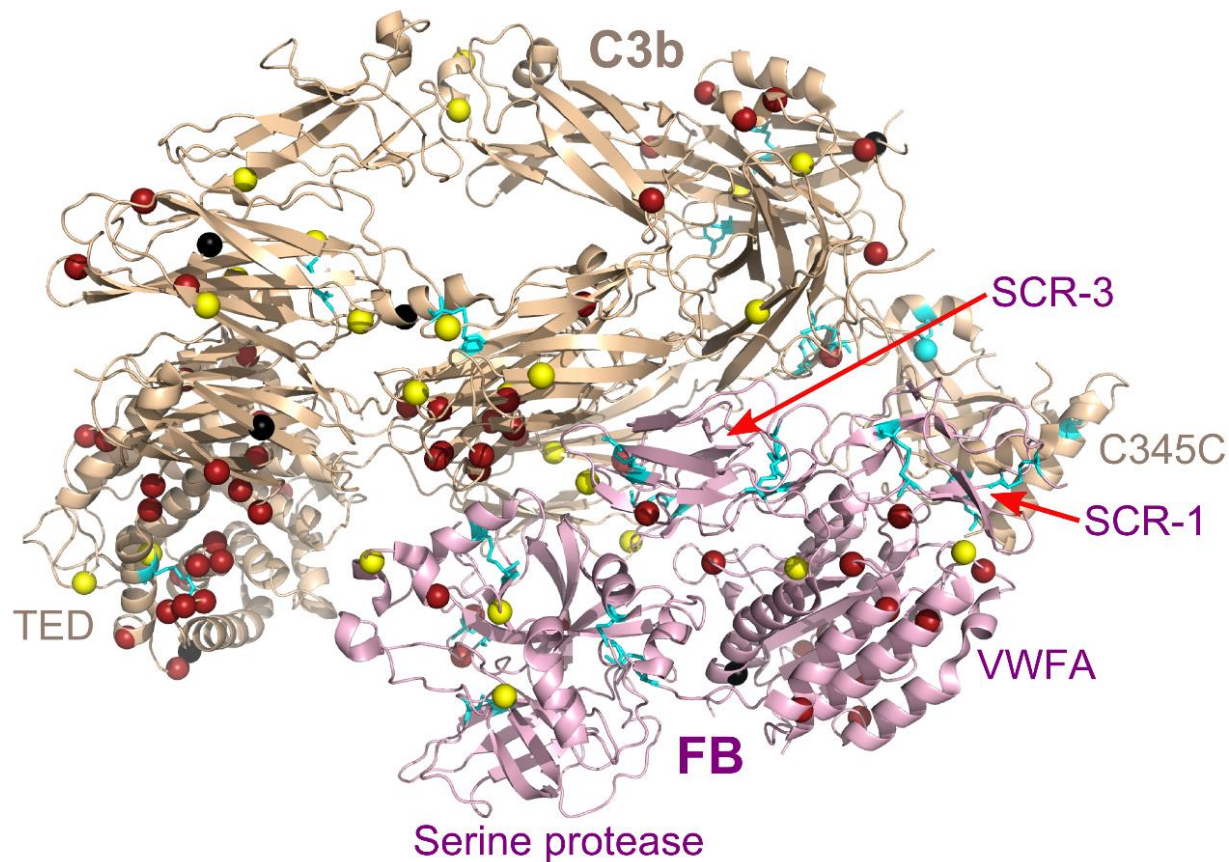


Figure 6.3 Rare missense variants mapped onto the crystal model for the C3 convertase. Rare missense variants for aHUS, C3G and both aHUS and C3G are mapped as red, yellow and black spheres, respectively, onto the structural model for the C3b and factor B convertase complex (PDB code: 2XWJ). For C3, there were 61 aHUS (56 unique residues) and 24 C3G (24 unique residues) rare missense variants, of which five in each dataset were common to both aHUS and C3G (five unique residues). For FB, there were 19 aHUS (19 unique residues) and seven (seven unique residues) C3G rare missense variants, with one variant common to both aHUS and C3G. The models are shown in cartoon style and the disulphide bonds are shown in cyan stick style by using PyMol. The single rare missense variants affecting a Cys residue is shown as a cyan sphere.

6.4.3 Structural locations of aHUS and C3G rare variants in FH

For the complex of N-terminal extended FH with C3b and FI (Figure 6.2B), the molecular views were quantified using the program PISA to assess the surface-accessibility of the FH residues (Table 6.3). Mature FH contains 1213 residues. For free FH, 42 out of the 1213 FH residues were predicted to be surface-inaccessible (4%), and 1171 were accessible. For the complex, overlaps were found between three residue types of (i) surface-accessible FH, (ii) surface-inaccessible FH, or (iii) C3b or FI binding interface FH (bracketed numbers, Table 6.3). For example, 126 of the 1171 surface-accessible residues (10%) were found to lie within the C3b and FI binding interfaces with a buried surface area greater than 0 Å². Of these 126 residues, 50 were involved in the binding interface with C3b (α-chain) (4%), 56 with C3b (β-chain) (5%), and 36 with FI (3%), with overlaps between them.

For both aHUS and C3G, the vast majority of rare missense variants affected surface-accessible residues in FH (92% and 94%, respectively) given its large size (third column, Table 6.3), and the FH-FI-C3b crystal structure will be uninformative as to how these leads to disease. However, in order to take the number of residues in the three types into account (top row Table 6.3), the proportion of residues for each residue type were now considered. For aHUS, surface-inaccessible residues within FH were most affected by aHUS (21%; 9 out of 42 surface-inaccessible residues) (last column, Table 6.3). This implies that a misfolded FH caused the phenotype. The 1171 FH residues that were surface-accessible were affected by aHUS at a frequency of 8% (third column, Table 6.3). Thus, the surface-inaccessible residue frequency (21%) was significantly greater than the surface-accessible frequency (8%; $p=0.0086$) only, the latter including the 7 and 6 variant residues involved with C3b. Altered FH residues at the interface can also contribute to disease. In contrast, for C3G, the type of residues most affected by C3G were the surface-accessible residues in the C3b binding interface (β-chain) (5%) (Table 6.3). However, this frequency was not significantly greater than for any of the other types, these being all similar in frequency for C3G. In addition, there were fewer C3G variants than aHUS variants in the dataset, which reduces the power of these results.

These analyses above were based on the numbers of aHUS and C3G affected residues only. In order to consider how many times each residue was mutated in aHUS and C3G, the analyses were repeated with allele count data for aHUS and C3G (Table

Table 6.3 FH residue surface-accessibility analyses for aHUS and C3G

Residue type	Total residues	Surface-accessible residues ^a	Surface-accessible binding interface residues ^b (unique)			Surface-inaccessible residues ^c
			C3b α	C3b β	FI	
FH all	1213	1171 (1045)	50 (42)	56 (50)	36 (34)	42
<i>of FH all (%)</i>	100	97	4	5	3	4
FH aHUS	107	98 (89)	7 (5)	6 (4)	0	9
<i>of FH aHUS (%)</i>	100	92	7	6	-	8
<i>of FH all (%)^d</i>	9	8	14	11	-	21
allele count	311	308 (290)	12 (8)	8 (4)	0	13
<i>of FH aHUS (%)</i>	100	99	4	3	-	4
<i>of FH all (%)</i>	26	26 (28)	24 (19)	11 (16)	-	31
FH C3G	17	16 (12)	1 (1)	3 (3)	1 (1)	1
<i>of FH C3G (%)</i>	100	94	6	18	6	6
<i>of FH all (%)</i>	1	1	2	5	3	2
allele count	27	26 (26)	2 (2)	3 (3)	2 (2)	1
<i>of FH C3G (%)</i>	100	96	7	11	7	4
<i>of FH all (%)</i>	2	2 (3)	4 (5)	5 (6)	6 (6)	2

^a Residues with accessible surface area greater than 0 Å², calculated by PISA, followed by the number of unique residues in brackets.

^b Solvent accessible residues with buried surface area greater than 0 Å², calculated by PISA.

^c Residues with accessible surface area of 0 Å², calculated by PISA.

^d The frequency of all FH residues for that residue type. For example, for the surface-accessible residues (third column), 98 out of 1171 were affected by aHUS, giving a frequency of 8%.

6.3). The 128 aHUS rare missense variants in FH mostly affected three types of residues: nine of the 42 surface-inaccessible (or buried) in 13 case alleles (31%), followed by 89 of the 1045 non-binding surface-accessible residues (290 case alleles; 28%) and finally nine of the 92 residues at the two C3b-binding interfaces (12 case alleles; 13%), but none at the interfaces with FI. In contrast, for 19 C3G variants, in addition to these three totalling 2% (one case allele), 3% (26 case alleles) and 5% (five case alleles) respectively, two case alleles (6%) within the FI binding interface were now involved.

For the 36 surface-accessible FH residues at the FI binding interface, none were affected by aHUS, but one was affected by C3G in the rare missense variant datasets (sixth column, Table 6.3). However, this one C3G residue corresponded to two C3G variant alleles: p.Arg127Cys and p.Arg127His (Table 6.3). In addition, an additional RV p.Arg127Leu was reported in the literature for C3G in two MPGN patients, both in homozygous form (four extra alleles in total) (Dragon-Durey et al., 2004). In the C3b-FH-FI complex, p.Arg127 was predicted by PISA software to form a salt bridge and a hydrogen bond with FI p.Asp438 (interatomic distance: 2.88 Å). For aHUS, for the FI binding interface, one variant p.Arg166Leu, of only one allele count, was also reported in the literature (Chaudhary et al., 2014). Overall, rare missense variants in FH that occurred in the FI binding interface appeared to be more frequent in C3G (either two or six alleles; 6% or 18%) than in aHUS (either zero or one allele; 0% or 3%), but more studies are required to verify this.

6.4.4 Frequency of aHUS and C3G rare variants in protein disulphides

In order to test that the disruption of disulphides was significant in causing aHUS and C3G, each of the aHUS, C3G and genomic reference ExAC datasets were used to calculate the AF of rare missense variation that affected disulphide bridge-forming Cys residues in each of FH, MCP, C3, FI and FB (Table 6.4). These type of variants are associated with a LoF in the protein.

For FH, all 80 Cys residues are involved in 40 essential disulphide bonds that stabilise the SCR domains (Figure 6.1). For FH in aHUS, the 21 affected Cys residues (24 variants; red and black text, Figure 6.1) were randomly distributed in FH at a total aHUS AF of 0.512% (Table 6.4). For FH in C3G, three rare missense variants affected Cys at a total AF of 0.451% and all of these three were also seen in the aHUS dataset

Table 6.4 The frequency of rare variation in Cys residues that affected protein disulphide bonds for aHUS, C3G and ExAC

Gene	aHUS				C3G				ExAC			
	Number of				Number of				Number of			
	rare missense variants (residues)	Allele count	Allele number	Allele frequency (%)	rare missense variants (residues)	Allele count	Allele number	Allele frequency (%)	rare missense variants (residues)	Allele count	Allele number	Allele frequency (%)
<i>CD46</i>	9 (8)	30	5884	0.510	0 (0)	0	812	0.000	0 (0)	0	116676	0.000
<i>CFB</i>	0 (0)	0	4914	0.000	0 (0)	0	758	0.000	1 (1)	1	117464	0.001
<i>CFH</i>	24 (21)	32	6256	0.512	3 (3)	4	886	0.451	9 (9)	10	118326	0.009
<i>C3</i>	0 (0)	0	4910	0.000	1 (1)	1	758	0.132	0 (0)	0	119090	0.000
<i>CFI</i>	5 (5)	6	5846	0.103	0 (0)	0	816	0.000	9 (9)	12	120578	0.010

(p.Cys431Tyr, p.Cys915Ser, p.Cys1218Arg) (black text, Figure 6.1). For FH in ExAC, only nine out of 118,326 total mean FH alleles (0.009%) harboured a Cys-affecting rare missense variant in FH (Table 6.4). The total mean alleles were calculated as previous for the RV burden calculations (Osborne et al., 2018a). These nine ExAC alleles corresponded to nine RVs and these were distributed in SCR-1, SCR-3, SCR-5, SCR-7, SCR-9 and SCR-19. For FH, the AF of Cys-affecting rare missense variation for both aHUS (0.512%) and C3G (0.451%) was significantly greater than for ExAC (0.009%; both $p < 0.0001$).

For MCP, rare missense variants in disulphide bridge-forming Cys residues were only seen for aHUS. None were seen for C3G. Thus, for MCP in aHUS, nine variants affected eight different Cys residues and these were all located in the four SCR domains (blue text, Figure 6.1). Their total aHUS AF of 0.510% was significantly greater than for ExAC (0%; $p < 0.0001$) (Table 6.4).

For FI, the total AF of rare missense variation in disulphide bridge-forming Cys residues was significantly greater for aHUS (0.103%) than for ExAC (0.010%; $p < 0.0001$) (Table 6.4). None were observed for C3G.

For C3, only the C3G dataset contained a rare missense variant that affected a disulphide bond-forming Cys residue. This variant, p.Cys1518Arg, occurred in the C345C domain (cyan sphere, Figure 6.3). The C3 Cys1518 residue is conserved in ten vertebrates (Database of Complement Gene Variants, AA Alignments webpage). When compared to ExAC (0%), the total AF of these C3G variants (0.132%) was significantly greater ($p < 0.0001$) (Table 6.4). This was not the case for aHUS.

For FB, the AF of rare missense variation in disulphide bridge-forming Cys residues was zero for both aHUS and C3G, and 0.001% for ExAC (Table 6.4).

6.5 Discussion

It is unknown whether the structural locations of rare missense variants in the complement proteins differ between aHUS and C3G. This is important for the prediction of aHUS or C3G in patients. In this chapter, for the structural locations of complement variants, a number of differences were seen between aHUS and C3G, which indicated

their potential use for the prediction of aHUS or C3G. Firstly, the aHUS rare missense variants appeared to cluster at either the regulatory FH SCR-3/4 and C3b MG2/CUB or TED/MG1, or cell surface SCR-20 associated regions, whereas the C3G variants appeared to cluster in regions associated with decay acceleration activity (SCR-1/2), some cofactor activity (SCR-1/3) and FI-mediated cleavage. Secondly, for rare missense variants in FH, those that occurred in the FI binding interface residues appeared to be more frequent in C3G (either two or six alleles; 6% or 18%) than in aHUS (either zero or one allele; 0% or 3%). However, more statistical based analyses are required to verify this. Thirdly, for Cys residues involved in disulphide bonds, only those in the regulators FH, MCP and FI were associated with aHUS, while only those in FH and, interestingly, C3 were associated with C3G, when compared to the ExAC reference dataset.

6.5.1 Consensus SCR domain updated for aHUS and C3G

The SCR consensus domain has proved useful in earlier analyses of genetic variants in complement ([Rodriguez et al., 2014](#); [Saunders et al., 2007](#)). Here, it was found useful to analyse the abundances of the aHUS and C3G RVs in the functional regions of complement SCR domains, following an update for FH and MCP ([Figure 6.1](#)). Overall, the four conserved Cys residues showed the highest normalised both allele count and AF of 65% and 0.0104%, respectively ([Table 6.2](#)). For the locations of residues in either the hypervariable loop, not in the hypervariable loop nor the Cys residues, or all locations, the aHUS AFs were very similar, ranging from 0.0039% to 0.0044% ([Table 6.2](#)). In summary, the four conserved Cys residues with a total aHUS AF of 0.0104% were more affected by aHUS rare missense variants than the rest of the consensus SCR domain. The hypervariable loop residues were not significantly more affected by aHUS than the rest of the consensus SCR domain residues. For aHUS, this outcome suggested that the hypervariable loop did not have greater functional importance when compared to the other residues of the consensus SCR domain which were not involved in disulphide bridges (Cys). Previously, the consensus SCR hypervariable loop was associated with disease variants for AMD, aHUS and C3G, for which the analyses included the AMD-risk variant p.His402Tyr ([Rodriguez et al., 2014](#)). For aHUS, p.Tyr402His (rs1061170) is part of the protective *CFH*_{CATAGG} haplotype for aHUS. For dense deposit disease, a sub-category of C3G, p.Tyr402His is a risk allele, and the presence of two or more risk alleles increased the odds ratio of developing the disease ([Abrera-Abeleda et al., 2011](#); [Smith et al., 2011](#)). Thus, by itself, p.His402Tyr does not predispose for either aHUS or C3G. The

analyses of common risk or protective haplotypes for aHUS and C3G was beyond the scope of this thesis.

6.5.2 aHUS variants cluster in regulatory FH-C3b regions

Three-dimensional views assist with the interpretations of the variants. Firstly, aHUS and C3G rare missense variants were mapped onto the complexes of FH-C3b-FI (Wu et al., 2009; Xue et al., 2017) and FH-C3dg (Kolodziejczyk et al., 2017) with improved structural models for glycosylated full length FH (Chapter 5), and the C3b-FB convertase (Forneris et al., 2010). Compared with previous analyses of their distributions in each of FH, C3, FI, MCP and FB separately (Chapter 4), this chapter focusses on the protein complexes with FH in its two major conformations (Chapter 5). This permitted a more detailed understanding of the variant distributions and their associations with the functional binding regions of FH. For aHUS and FH, the rare missense variants were clustered at SCR-20 and distributed non-randomly throughout the other SCR domains (red and black spheres, Figures 6.2A, B and C), as previously observed at the genetic level (Chapter 4, Section 4.5.4). In addition, for the FH-C3b complex, clusters of aHUS variants were observed at the interfaces between FH SCR-3 and C3b MG2/CUB, and also between FH SCR-4 and C3b TED/MG1 (red and black spheres, Figure 6.2A and B). This indicated the association of aHUS with disrupted FH-C3b binding, which leads to a LoF for FH. For the FH SCR-4 and TED/MG1 interface, this region had previously been associated with individual aHUS variants e.g. C3 p.Gln1139Lys in TED introduces a positive charge that may lead to formation of a salt bridge between p.Lys1139 and the preceding p.Asp1138, thereby modifying the C3b-FH interaction (Wu et al., 2009). For both SCR-3 and SCR-4, their binding to residues within the C3b domains MG1, MG2, CUB and TED holds C3b in place for cleavage by FI. Thus, both SCR-3 and SCR-4 are important for the regulation of C3b by FH cofactor activity and decay acceleration activity. Overall, for FH and C3b, the aHUS rare missense variants appeared to be clustered at both the regulatory (SCR-3 and C3b MG2/CUB, and SCR-4 and C3b TED/MG1) and cell surface (SCR-20) associated regions.

For C3G, the rare missense variants were distributed mainly in C3 and in FH SCR-1/3 (yellow and black spheres, Figure 6.2A, B and C). FH SCR-1/2 are required for displacing Bb in decay acceleration activity. Interestingly, in contrast to aHUS, there were no C3G rare missense variants in SCR-4 or the interfaces between SCR-4 and TED or

MG1. This implied that C3G is not associated with the disruption of C3b regulation by FH via these SCR-4 interfaces. Few C3G rare missense variants were seen in the middle and C-terminal SCR domains. By this, the C3G variants appeared to cluster in regions associated with decay acceleration activity (SCR-1/2) and some cofactor activity (SCR-1/3) only, and not in regions required for cell surface regulation (SCR-19/20) ([Chapter 4, Section 4.5.4](#)). As discussed in [Chapter 4](#), this reflects the role of fluid-phase and surface-associated activities in the C3G and aHUS phenotypes. For C3b, two C3G variants were located in the first and second cleavage sites thus prevent cleavage of C3 by FI and a cofactor. These represent GoF variants in C3 that may lead to an over-active C3 and thus its consumption. In contrast, in aHUS, none of these C3 cleavage sites were affected.

6.5.3 Structural locations of FH variants differ between aHUS and C3G

For a protein, the change of a surface-associated residue involved in a binding interface generally affects the activity of the protein only, and not its amounts or levels. On the other hand, buried residues are involved in structural stabilisation, and their disruptions via missense variants can lead to reduced levels of the protein due to protein misfolding. Thus, the locations of rare missense variants can account for their effects on protein structure, function and the associated biochemical pathway. For aHUS and C3G, their occurrences in individuals may be predicted from complement genetic variant data, by analysing the structural locations of the complement variants. For N-terminal extended FH in complex with C3b and FI ([Figure 6.2B](#)), the program PISA was used to assess the surface-accessibility of the FH residues ([Table 6.3](#)). By this, 97% FH residues were surface-associated and of these, 4%, 5% and 3% were involved in molecular interactions with C3b (α -chain), C3b (β -chain) and FI, at SCR-1/4, respectively. In addition, 4% residues were buried, and were thus predicted to be involved in structural stabilisation for FH. For the FH residues involved in interactions, only those that interact with C3b and FI at the N-terminal SCR-1/4 were analysed here. For FH, the 128 aHUS rare missense variants mostly affected three types of residues; nine of the 42 surface-inaccessible (or buried) residues in 13 case alleles (31%), followed by 89 of the 1045 non-binding surface-accessible residues (290 case alleles; 28%) and finally nine of the 92 residues at the C3b-binding interfaces (12 case alleles; 13%). None of the residues at the FI binding interfaces were affected. In contrast, for the 19 C3G variants, in addition to these three types totalling 2% (one case allele), 3% (26 case alleles) and 5% (five case alleles) respectively, two case alleles (6%) within the FI binding interface were now involved. For the C3G

dataset of rare missense variants analysed for this chapter, this corresponded to two alleles for p.Arg127Cys and p.Arg127His. However, in the literature, a further four alleles for C3G were reported for the variant p.Arg127Leu ([Dragon-Durey et al., 2004](#)), and this may increase the frequency to 18%. In the C3b-FH-FI complex, p.Arg127 was predicted by PISA software to form a salt bridge and a hydrogen bond with FI p.Asp438 (interatomic distance: 2.88 Å). Thus, in the C3b-FH-FI complex, mutation of the FH Arg127 residue may lead to decreased FI binding and cofactor activity. This may lead to an over-active complement phenotype, which is usually observed for C3G. For the aHUS dataset analysed for this thesis chapter, no alleles for the FI binding region of FH were reported. However, one aHUS variant (p.Arg166Leu) in this region was reported in the literature, for one aHUS allele ([Chaudhary et al., 2014](#)). Overall, rare missense variants in FH that occurred in the FI binding interface appeared to be more frequent in C3G (either two or six alleles; 6% or 18%) than in aHUS (either zero or one allele; 0% or 3%), however more statistical analyses are required to verify this.

6.5.4 Disulphide bond disruption is associated with aHUS and C3G

For proteins, the disruption of a structural disulphide bond often leads to LoF via protein misfolding and lower levels of the protein in plasma. For the SCR domains in FH, MCP and FB, all the Cys residues were involved in disulphide bridges. Thus, for FH, 80 Cys residues are involved in 40 essential disulphide bonds that stabilise the SCR domains ([Figure 6.1](#)).

For FH, the frequency of Cys-affecting rare missense variation for both aHUS (0.512%) and C3G (0.451%) was significantly greater than for ExAC (0.009%; both $p < 0.0001$) ([Table 6.4](#)). For aHUS, the 21 affected Cys residues were randomly distributed in FH. For C3G, the three affected Cys FH residues were also affected by aHUS (residues 431, 915 and 1218) which suggested that the structural defect of disulphide bridge disruption in FH in SCR-7, SCR-15 and SCR-20 was common to both the aHUS and C3G phenotypes. For FH, such disulphide bond disruption would lead to misfolding of the SCR domain, which causes poor secretion or a short half-life ([Wagner et al., 2016](#)). Interestingly, for ExAC, nine FH alleles harboured a Cys-affecting missense rare variant, however the overall frequency of these for the dataset (0.009%) was very low. For these ExAC variants, the individuals harbouring them must possess a combination of genetic or environmental factors that had so far protected them from aHUS or C3G, which lowers

their penetrance ([Chen et al., 2016](#)). An example of such genetic factors is the complotype ([Heurich et al., 2011](#)).

For C3, ten structural disulphide bonds were present in its most recent structure ([Xue et al., 2017](#)). Of interest for C3, only the C3G dataset contained rare missense variation that affected disulphide bond-forming Cys residues. This corresponded to the variant p.Cys1518Arg, which was predicted to disrupt the disulphide bond with p.Cys1590 in the C345C domain. The variant p.Cys1518Arg was previously found in a patient with severe C3 deficiency and predicted to cause a LoF for C3. This may lead to either the retention of C3 in the ER or C3 with a misfolded C345C domain. However, C3G is associated with over-activation of complement and therefore a GoF phenotype in the activator C3. Thus, either p.Cys1518Arg is not truly associated with C3G, or the breaking of the disulphide bridge with p.Cys1590 is beneficial for a catalytic or allosteric C3 event which enhances C3 function. This may be related to the formation of the C3 convertase which is mediated by C3 C345C binding to FB ([Forneris et al., 2010](#)). By this, despite that p.Cys1518Arg does not map to the C3b-FB binding site, it may enable greater flexibility within the C345C domain for FB binding ([Figure 6.3](#)). For the C345C domain, this may indicate a novel structural mechanism that enables C3b-FB binding for the formation of C3 convertase complexes. aHUS was not associated with disulphide bridge disruption in C3 ([Table 6.4](#)), as expected.

Chapter Seven

Conclusions – new findings on complement genetics and protein structures

7.1 Prologue

7.1.1 Overview of complement function and the role of FH

Complement is an evolutionarily ancient inflammatory pathway that provides rapid protection from pathogens. The AP of complement is spontaneously activated at low levels in order to provide a tick-over mechanism of protection. Due to the broad specificity of activated C3b for any nucleophilic surface, host cell surfaces require continuous protection from complement activation, whilst pathogens remain targeted. In healthy individuals, this is achieved by a fine balance between the activator and regulator components which includes host cell-bound regulators ([Chapter 1, Section 1.6](#)). For the AP, factor H (FH) is the main inhibitor of C3 activation ([Rodriguez de Cordoba et al., 2004](#)) and works with or alongside the two other regulators FI and MCP to regulate the activators FB and properdin. Specifically, FH regulates the AP by providing cofactor activity for the FI-mediated degradation of C3b, DAA for the dissociation of the C3 and C5 convertases and by blocking the formation of the C3 convertases. In order to provide such regulation at host cell surfaces, FH recognises surface poly-anions such as sialic acid and glycosaminoglycans. The N-terminal SCR-1/4 domains interact with C3b and FI for decay-accelerating and cofactor activity. The last two C-terminal domains, SCR-19 and SCR-20, interact with the C3b TED and the host surface thereby mediating self-recognition and protection. This self-cell specificity depends on the simultaneous binding of SCR-19/20 to host cell surface polyanionic markers and C3b ([Schmidt et al., 2013](#)). In addition, the FHR 1-5 proteins support FH cofactor activity as well as additional complement regulatory activities, but can also compete with FH for C3b binding ([Chapter 1, Section 1.6.1](#)).

7.1.2 Complement genetic variants in disease

Genetic variants in the complement AP can lead to a range of disease phenotypes, from ultra-rare, sudden and severe, to more common, chronic and progressive. Such variants affect the levels and/or the biochemical function of the complement proteins. For genetic variants that lead to changes in the protein, their locations in different types of functional sites will typically have different consequences. For example, residues involved in molecular interactions are mainly surface-associated whereas residues involved in structural stabilisation are typically buried or surface-associated ([Chapter 2,](#)

[Section 2.4](#)). For genetic diseases, characterisation of the genotype-phenotype relationship allows predisposing, risk and protective factors to be identified. When the structure of a protein is known, the impact of the genetic variant on protein function and the genotype-phenotype relationship can be much better understood. For patients with genetic disease, this can be essential for predicting the disease outcome, prescribing treatments and developing new therapies.

aHUS and C3G are two ultra-rare severe diseases that are associated with dysregulation and over-activation of the complement AP. In aHUS, an immune insult triggers complement activation, primarily by the AP, which cannot be properly controlled due to either underlying genetic defects (~60% aHUS cases) or acquired factors such as FH auto-antibodies (5-13% aHUS cases) ([Chapter 2, Section 2.11](#)). This leads to endothelial cell attack and microvasculature occlusion ([Caprioli et al., 2006](#)), which results in organ damage to the kidneys, gastrointestinal tract, liver, pancreas and brain. In C3G, uncontrolled complement activation, from either underlying genetic complement variants (~20%) or acquired auto-antibodies such as C3Nef (~80%), leads to deposits of C3 and its fragments within the kidney glomerulus and its damage ([Chapter 2, Section 2.11](#)). In contrast, AMD is a common, progressive eye disease which affects ~50 million people across the world and is the leading cause of blindness in developed countries ([Clark et al., 2014](#)). In AMD, polymorphous debris known as drusen are deposited between the RPE and Bruch's membrane. Similar to aHUS and C3G, complement genetic variants are thought to increase AMD risk by affecting the functionalities or circulating levels of the complement system, such as FH ([Chapter 1, Section 1.7.3](#)).

For patients with genetic disease, new genetic variants are continuously being identified due to the onset of more rapid next generation sequencing and advanced computing methods. The current challenge for researchers of genetic diseases is to separate disease-associated genetic variants from the broader background of variants present in all human genomes that are rare, potentially functional, but not pathogenic for the disease ([Vieira-Martins et al., 2016](#)). This was one major question addressed in this thesis ([Chapters 4, 5 and 6](#)). For each variant, this may be achieved by using a combination of familial segregation analyses, statistical analyses, which are based on the frequencies of variants in disease and reference populations, functional assays, structural biology and *in silico* predictive analyses. Background variation can be well represented, with caveats, by reference datasets sourced from large sequencing projects such as the

ExAC (Lek et al., 2016), the 1000GP (Genomes Project et al., 2010) and the EVS (Fu et al., 2013; Tennessen et al., 2012). For genetic variants, as their (allele) frequency in the general (reference) population increases, their pathogenicity generally decreases (Kobayashi et al., 2017) (Section 2.4.3). Thus, a variant expected to be causative for a rare disease, such as aHUS or C3G, will not be frequent in unselected individuals without the disease. On the other hand, disease-risk alleles that are common in the general population, such as *CFH* p.Tyr402His, often show minimal effects on overall reproductive fitness. Common diseases such as AMD are often complex, and are thought to comprise either many common variants of small effect sizes which are not fully identified by underpowered GWAS (“common disease, common variant”), rare variants (“common disease, rare variant”) (Wagner, 2013) and/or further environmental factors. Overall, AF data is essential for assessing variant pathogenicity, especially for rare diseases (Sections 2.8.2, 2.8.3 and 2.9), and can be used to statistically verify disease-gene associations such as for statistical burden testing (Section 2.10).

Prior to the work presented in this PhD thesis, in order to analyse genetic variants identified in aHUS and other complement diseases, an interactive FH-HUS web-database (<http://www.fh-hus.org>) was set up at University College London in 2006 for both clinicians and researchers (Saunders et al., 2007; Saunders et al., 2006; Saunders & Perkins, 2006) (<http://www.fh-hus.org>). By 2014, there were 193 *CFH*, 130 *CFI*, 86 *CD46* and 64 *C3* variants for aHUS and other complement diseases (Rodriguez et al., 2014). However, major omissions from the 2006 FH-HUS database included not only new complement genes and variants associated with aHUS and related complement diseases, but also the inclusion of variant AF data. Thus, I created a new Database of Complement Gene Variants (Chapter 4) which incorporated these new variant and AF data, and used this to verify the association of RVs in the complement and related genes with aHUS and C3G.

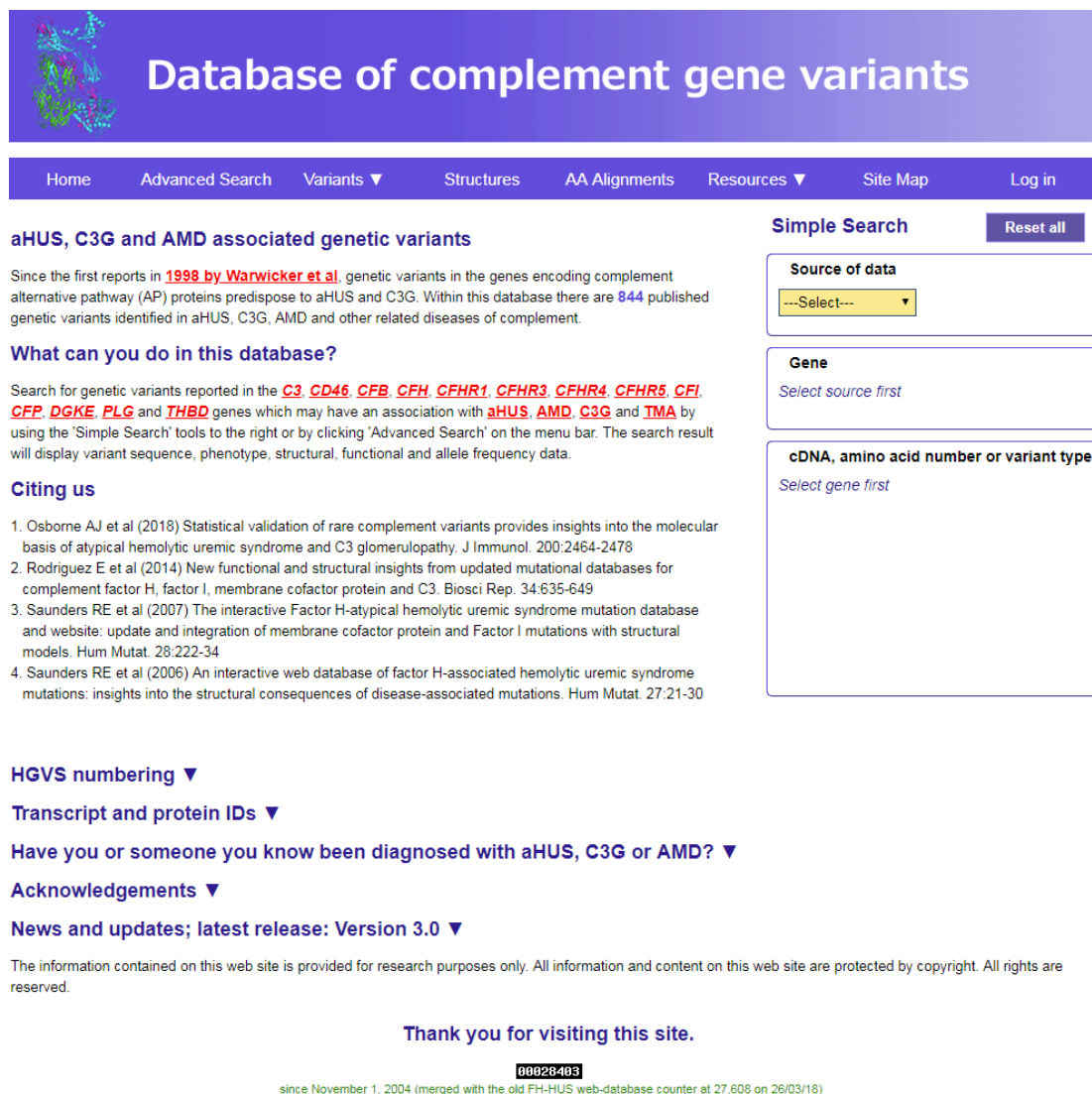
7.2 Statistical validation of rare complement variants provides insights on the molecular basis of atypical haemolytic uraemic syndrome and C3 glomerulopathy

aHUS and C3G are two ultra-rare severe diseases that are associated with dysregulation and over-activation of the complement alternative pathway. Typically, gene analysis for aHUS and C3G is undertaken in small patient numbers, yet it is unclear which genes most frequently predispose to aHUS or C3G. Accordingly, in this PhD thesis

(Chapter 4), I performed a six-centre analysis of 610 rare genetic variants in 13 mostly complement genes (*CFH*, *CFI*, *CD46*, *C3*, *CFB*, *CFHR1*, *CFHR3*, *CFHR4*, *CFHR5*, *CFP*, *PLG*, *DGKE*, and *THBD*) from >3500 patients with aHUS and C3G. This reported 371 novel rare variants for aHUS and 82 for C3G. My new interactive Database of Complement Gene Variants (<http://www.complement-db.org>) was used to extract allele frequency data for these 13 genes using the ExAC server as the reference genome. For aHUS, significantly more protein-altering rare variation was found in five genes *CFH*, *CFI*, *CD46*, *C3* and *DGKE* than in ExAC (allele frequency <0.01%), thus correlating these with aHUS. For C3G, an association was only found for rare variants in *C3* and the N-terminal C3b-binding or C-terminal non-surface-associated regions of *CFH*. In conclusion, the RV analyses showed non-random distributions over the affected proteins, and different distributions were observed between aHUS and C3G that clarify their phenotypes. Importantly, these results were able to explain how changes in the same biochemical pathway and/or proteins result in the different pathologies observed with aHUS and C3G, which was not clear before. Furthermore, this work has reduced the earlier knowledge gap in the genetics and genotype-phenotype correlations of C3G to bring these closer to that of aHUS (Goodship et al., 2017).

7.2.1 The Database of Complement Gene Variants and future work

In summary, the new Database of Complement Gene Variants (<http://www.complement-db.org>) enhances understanding of rare genetic variants in aHUS and C3G for clinical applications (Figure 7.1). Improvements include the use of an ExAC AF cut-off in the search tools (Figure 7.2), the display of AF data for the disease and references datasets (Figure 7.3), predictive comparisons of wild-type and mutant amino acids, *in silico* analyses using PolyPhen-2 and SIFT (Figure 7.4), examination of evolution-conserved residues across species, and correlations with functional binding sites. These tools enable clinicians to assess RVs in disease, for example, to investigate which variants within these genes conferred predisposition to aHUS and C3G, and to identify mutational hotspots within these protein structures. This can be especially useful for variants of uncertain significance for which no experimental data currently exists. Through the use of AFs and burden testing in the database, the new database informs patient management by enabling clinical immunologists to interpret new variants in terms of their associations with aHUS and C3G (Chapter 4).



Database of complement gene variants

Home Advanced Search Variants ▼ Structures AA Alignments Resources ▼ Site Map Log in

aHUS, C3G and AMD associated genetic variants

Since the first reports in [1998 by Warwicker et al.](#), genetic variants in the genes encoding complement alternative pathway (AP) proteins predispose to aHUS and C3G. Within this database there are **844** published genetic variants identified in aHUS, C3G, AMD and other related diseases of complement.

What can you do in this database?

Search for genetic variants reported in the [C3](#), [CD46](#), [CFB](#), [CFH](#), [CFHR1](#), [CFHR3](#), [CFHR4](#), [CFHR5](#), [CFI](#), [CFP](#), [DGKE](#), [PLG](#) and [THBD](#) genes which may have an association with **aHUS**, **AMD**, **C3G** and **TMA** by using the 'Simple Search' tools to the right or by clicking 'Advanced Search' on the menu bar. The search result will display variant sequence, phenotype, structural, functional and allele frequency data.

Citing us

- Osborne AJ et al (2018) Statistical validation of rare complement variants provides insights into the molecular basis of atypical hemolytic uremic syndrome and C3 glomerulopathy. *J Immunol.* 200:2464-2478
- Rodriguez E et al (2014) New functional and structural insights from updated mutational databases for complement factor H, factor I, membrane cofactor protein and C3. *Biosci Rep.* 34:635-649
- Saunders RE et al (2007) The interactive Factor H-atypical hemolytic uremic syndrome mutation database and website: update and integration of membrane cofactor protein and Factor I mutations with structural models. *Hum Mutat.* 28:222-34
- Saunders RE et al (2006) An interactive web database of factor H-associated hemolytic uremic syndrome mutations: insights into the structural consequences of disease-associated mutations. *Hum Mutat.* 27:21-30

Simple Search **Reset all**

Source of data
---Select---

Gene
Select source first

cDNA, amino acid number or variant type
Select gene first

HGVS numbering ▼

Transcript and protein IDs ▼

Have you or someone you know been diagnosed with aHUS, C3G or AMD? ▼

Acknowledgements ▼

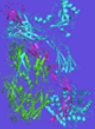
News and updates; latest release: Version 3.0 ▼

The information contained on this web site is provided for research purposes only. All information and content on this web site are protected by copyright. All rights are reserved.

Thank you for visiting this site.

00026403
since November 1, 2004 (merged with the old FH-HUS web-database counter at 27,608 on 26/03/18)

Figure 7.1 Screenshot of the Database of Complement Gene Variants homepage. The design is based on the European Association for Haemophilia and Allied Disorders coagulation Factor IX web-database (<http://www.factorix.org/>) (Rallapalli et al., 2013).



Database of complement gene variants

[Home](#)
[Advanced Search](#)
[Variants ▼](#)
[Structures](#)
[AA Alignments](#)
[Resources ▼](#)
[Site Map](#)
[Log in](#)

[DISCLAIMER](#): Data from this Database may be downloaded only for personal research use. No part of the data in this Database may be presented or published elsewhere in any format, printed or electronic, without specific written permission from the Database Curators.

Advanced Search

- Genetic variation, phenotype and control data can be retrieved from this database using any combination of the following search criteria.
- If nothing is selected for a search category, all results of the category will be retrieved from the database.
- The ExAC allele frequency cut-off default value is 0.01 (1%). Please enter a different value if required.

Search

Reset

Source of data

---Select---

☐ Search for variants that are exclusively from this source.

Gene

Select source first

Domain, Location and Residue

Select gene first

Reference

---Select---

☐ aHUS
 ☐ AMD
 ☐ C3G
 ☐ Other
 ☐ TMA
 ☐ Unknown

Condition

☐ Search for variants that are exclusively from cases with this condition.

No. of variants per case

☐ 1
 ☐ 2
 ☐ 3
 ☐ 4
 ☐ 5
 ☐ 6
 ☐ 7

The ExAC allele frequency cut-off value:

☐ Include genomic rearrangements and gross deletions not in ExAC

Search

Reset

Structural Immunology Group, University College London, Gower Street, London WC1E 6BT

Figure 7.2 Screenshot of the Advanced Search tools in the Database of Complement Gene Variants. For searching the database for variant data, the search terms can now include the type of condition, the number of variants per case and the Exome Aggregation Consortium (ExAC) allele frequency cut-off value for filtering based on a reference allele frequency.

CFH (FH)

c.1949G>T

p.Gly650Val (632)

The 1000 Genomes Project AF: 0.00040

Location: Exon(13)

Variant Type: Point

Variant Effect: Missense

Domain: SCR11

Transcript: ENST00000367429

Genomic: 1: 196695675 [GRCh37]

Phenotype ▶

Number	FH level	C3 level	FI level	MCP level	AntiFH Ab	Zygosity	Disease inheritance	Other variants	Condition	Reference
1265						Heterozygous	Sporadic		C3G	Osborne et al, 2018
1735	99	990	130	LOW	N	Heterozygous	Sporadic	CD46: c.608T>C (p.Ile203Thr)	aHUS	Osborne et al, 2018

Allele Frequency ▶

aHUS population (based on Osborne et al, 2018 only)

Allele Count (AC)	No. of patients screened for CFH	Allele Number (AN)	Allele Frequency (AF) estimate
1	3128	6256	0.00016

C3G population (based on Osborne et al, 2018 only)

Allele Count (AC)	No. of patients screened for CFH	Allele Number (AN)	Allele Frequency (AF) estimate
1	443	886	0.001129

Reference populations

dbSNP: rs143237092

ClinVar entry

The 1000 Genomes Project (1000GP):

1000GP Alleles	1000GP AC	1000GP AN	1000GP AF	P (aHUS)	P (C3G)
G/T (2)	2	5008	0.000399	0.1 < P < 0.4875	0.1 < P < 0.4875

ExAC:

ExAC Ethnicity	ExAC AC	ExAC AN	ExAC AF	P (aHUS)	P (C3G)
ALL	28	120926	0.000232	0.1 < P < 0.4875	0.1 < P < 0.4875
African	0	10368	0	0.1 < P < 0.4875	0.05 < P < 0.1
European	21	66482	0.000316	0.1 < P < 0.4875	0.1 < P < 0.4875
East Asian	0	8574	0	0.1 < P < 0.4875	0.05 < P < 0.1
Finnish	0	6592	0	0.1 < P < 0.4875	0.1 < P < 0.4875
Latino	6	11510	0.000521	0.1 < P < 0.4875	> 0.4975
Other	1	902	0.001109	0.1 < P < 0.4875	0.1 < P < 0.4875
South Asian	0	16498	0	0.1 < P < 0.4875	0.0125 < P < 0.025

Exome Variant Server (EVS):

EVS Ethnicity	EVS Alleles	EVS Allele	EVS AC	EVS AN	EVS AF	P (aHUS)	P (C3G)
ALL	G/T (2)	T	3	13006	0.000231	0.1 < P < 0.4875	0.1 < P < 0.4875
European-American	G/T (2)	T	3	8600	0.000349	0.1 < P < 0.4875	0.1 < P < 0.4875
African-American	G/T (2)	T	0	4406	0	0.1 < P < 0.4875	0.1 < P < 0.4875

Structure and Function ▶

No functional or structural studies for this variant (c.1949G>T) currently exist in the database.

Please click [here](#) to view the mapping of this variant onto the latest [FH](#) structure and view the functional analysis by [PolyPhen-2](#) and [SIFT](#).

Variant classification

(using guidelines in [ACMG](#) and [Goodship et al, 2017](#)): Likely benign

Figure 7.3 Screenshot of the Search Results web-page in the Database of Complement Gene Variants. For the complement factor H (CFH) variant, c.1949G>T p.650Val, its allele frequency for each of atypical haemolytic uraemic syndrome (aHUS), C3 glomerulopathy (C3G), The 1000 Genomes Project (1000GP), Exome Aggregation Consortium (ExAC) and Exome Variant Server (EVS) are displayed. For the ExAC, these allele frequencies are further categorised by ethnicity. Also provided are links to the variant in the databases of single nucleotide polymorphisms (dbSNP) and ClinVar.

In Depth Variation Analysis: c.1949G>T (p.Gly650Val)

CFH (FH) c.1949G>T p.Gly650Val (632) The 1000 Genomes Project AF: 0.00040

Location: Exon(13)

Domain: SCR11

Variant Type: Point

Transcript: [ENST00000367429](#)

Variant Effect: Missense

Genomic: 196695675 [\[GRCh37\]](#)

Residue Information:

	Name	Size	Charge	Hydrophobicity	Preferred position	Type	Class
Wild Type	Gly	small	neutral	hydrophobic	surface	-	-
Mutated	Val	medium	neutral	hydrophobic	buried	-	aliphatic

Structural Implications:

- Gly650 is shown in the structural model below as a red sphere.
- Gly650 is not in a proposed binding site.

Tool	Prediction	Binding site	PPH2	DSSP (SecStr)
PolyPhen-2 : HumDiv	benign	-	neutral	-
SIFT	Tolerated (score: 0.21)	-	-	-
Provean	Neutral (score: -2.14)	-	-	-

Java security warning? Visit the Help: Java Blocked page at www.java.com

Available structures:

- 3N0J**: Full length FH (original: [3N0J.pdb](#))

The default structure is shown below.

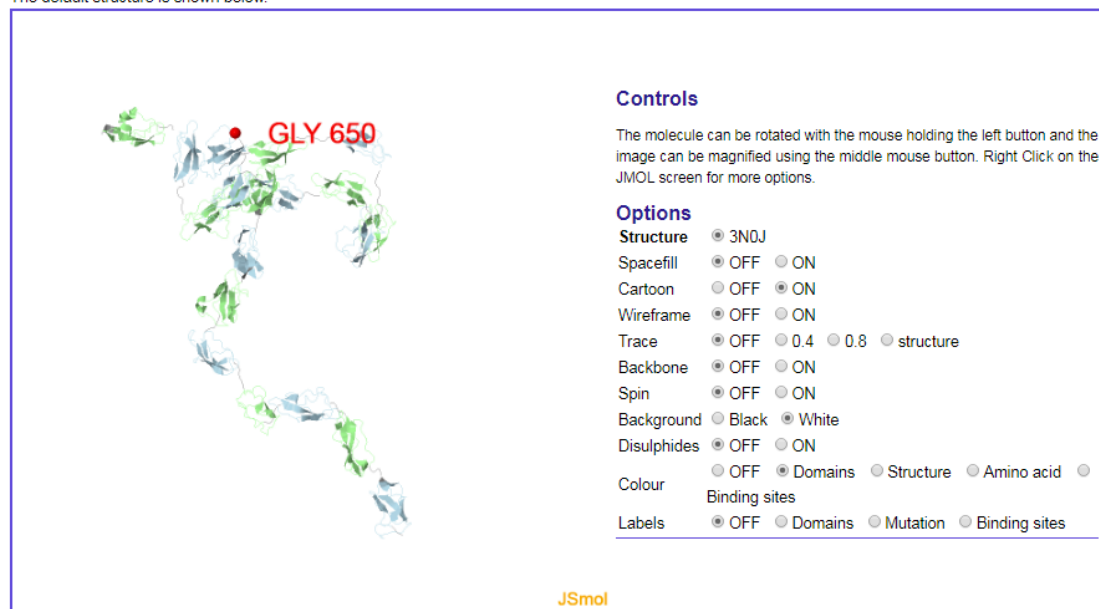


Figure 7.4 Screenshot of the In-Depth Variant Analysis web-page in the Database of Complement Gene Variants. In order to assess the structural and functional implications of a genetic variant in the database, this web-page does the following: the wild-type and mutated residue properties are compared, results from the *in silico* predictive tools PolyPhen-2, SIFT and Provean are displayed, and the variant is mapped onto the protein structure. This is shown for the complement factor H (CFH) variant, c.1949G>T p.650Val.

For the maintenance of the database in the long term, each collaborating group will provide an annual data update as a spreadsheet that will be automatically uploaded to the database with automated duplication and error checks that are programmed into the MySQL backend. As an open resource, other clinical centres are invited to upload formatted data updates through the curator. This will allow the continuing addition of new variants seen in the 13 complement and related genes of the original study ([Chapter 4](#)).

In addition, these clinical data can be extended to include other diseases of the complement system, most notably AMD, in order to better predict the outcome of disease variants. For example, a new sequencing study this year based on 866 aHUS/C3G and 697 AMD patients from the Nijmegen clinical centre ([Geerlings et al., 2018](#)) shows that the extension to AMD patients is feasible.

Reference genomic data, such as the ExAC, provide the baseline for assessing the significance of the genetic variants by referencing these to normal populations of variants. In the future, new variant data from the UK's ongoing 100,000 Genomes Project, for both reference and renal disease datasets, may be incorporated into the database in order to monitor the statistical associations of variants with aHUS and C3G. These analyses may also be extended to other complement diseases such as AMD.

For clinicians, in order to improve the understanding of the molecular basis of the complement protein defects, more accurate computational predictive tools that incorporate molecular dynamics simulations can be developed. This will provide clinicians with a rapid, informative and explorative view of the effect of missense variants on the complement proteins, and output the result in terms of a loss or a gain of function prediction.

7.3 Two distinct conformations of factor H regulate discrete complement-binding functions in the fluid phase and at cell surfaces

FH is the major regulator of C3b in the alternative pathway of the complement system in immunity. FH comprises 20 SCR domains, including eight glycans, and its Tyr402His polymorphism predisposes those who carry it for AMD. Biologically, given that the plasma concentration of FH is around 0.7 mg/ml ([Perkins et al., 2010a](#)), and AUC had demonstrated that FH forms oligomers ([Nan et al., 2008a](#)), reversible dimers and

trimers of FH are expected to form *in vivo* (Osborne et al., 2018b). The resulting clustering of FH on cell surfaces, when bound to surface markers and C3b/C3d, is likely to enable broader surface protection, thus being functionally significant (Ferreira et al., 2010). However, the accumulation of FH by clustering may also relate to the formation of soft drusen in AMD via chronic inflammation. In this PhD thesis, to better understand FH complement binding and self-association, the atomistic solution structures of both the His402 and Tyr402 FH allotypes were determined (Chapter 5).

Previously, in prior experimental work (Osborne et al., 2018b), analytical ultracentrifugation revealed that up to 12% of both FH allotypes self-associate, and this was confirmed by SAXS, mass spectrometry and SPR analyses. SAXS showed that monomeric FH has a radius of gyration R_g of 7.2-7.8 nm and a length of 25 nm. Thus, for this thesis, in order to model the monomeric FH structure, it was necessary for me to firstly extrapolate the scattering curves to zero concentration (Chapter 5).

Prior to this PhD thesis, preliminary molecular modelling of SAXS curves using an older technology based on conformational randomisation of the inter-SCR linkers revealed that full-length FH in solution has a folded-back SCR structure that is affected by ionic strength (Aslam & Perkins, 2001; Nan et al., 2010; Okemefuna et al., 2009).

For my work for this PhD thesis, starting from known structures for 17 SCR domains and eight glycans, the SAXS data were fitted using Monte Carlo methods to generate physically-realistic atomistic structures for monomeric FH (Chapter 5). Compared to previous, the FH modelling here is much improved by starting from 17 high resolution SCR structures (Figure 5.1; Table 5.1) (Hocking et al., 2008; Okemefuna et al., 2009; Schmidt et al., 2010; Wu et al., 2009) and utilising a much improved Monte Carlo and energy minimisation atomistic modelling method for the full-length FH structure (Perkins et al., 2016), including the eight FH glycan chains. The analysis of 29,715 physically realistic but randomised FH conformations resulted in 100 similar best-fit FH structures for each allotype. Two distinct molecular structures resulted that showed either an extended N-terminal domain arrangement with a folded-back C-terminus, or an extended C-terminus and folded-back N-terminus. These two molecular structures are the most accurate to date for glycosylated full-length FH. To clarify FH functional roles in host protection, crystal structures for the FH complexes with C3b and C3dg suggested that the extended N-terminal conformation accounted for C3b fluid phase regulation, the

extended C-terminal conformation accounted for C3d binding, and both conformations accounted for bivalent FH binding to anionic glycosaminoglycans on the target cell surface ([Chapter 5, Figure 5.8](#)). By this, FH with an N-terminal extended conformation appeared to be able to both dismantle fluid phase C3 convertases and perform cofactor activity. On the other hand, FH with a distinct C-terminal extended conformation will have increased affinity for binding to cell surfaces via both heparin (SCR-20) and deposited C3d (SCR-19) for enhanced local protection from complement activation. This is because the presence of C3d on the cell surface may increase the affinity of FH in the correct conformation for cell surface binding ([Kajander et al., 2011](#)). For this C-terminal extended FH structure, following its binding to cell surfaces, a conformational change involving the flexible linkers between SCR-4/5 and SCR-5/6 would allow SCR-1/4 to bind C3b. By this, FH attached to host cell surfaces can then also perform either cofactor activity (inactivation of C3b to iC3b) or DAA (separate C3b from Bb in the C3 convertase complex). Overall, by elucidating the conformational states of full-length FH and its structural interactions with C3b, FI and C3dg, the molecular basis for inflammatory processes has become better understood. This is relevant for the interpretation of complement genetic variants identified in inflammatory diseases such as AMD, and also aHUS and C3G. In addition, for the AMD-risk p.Tyr402His variant, the two FH Tyr402 and His402 allotypes displayed the same solution structures, thus the overall domain arrangement of FH was unaffected ([Chapter 5, Section 5.4.3](#)). This result is consistent with its common AF, and also the common frequency of AMD, for which only a subtle effect is expected.

7.3.1 Stoichiometry of FH and C3b complexes and future FH work

For the modelling of FH undertaken for this PhD thesis ([Chapter 5](#)), both of the resulting full-length solution-state FH models showed that SCR-1/4 and SCR-19 were too far apart to bind the same C3b molecule. This section ([7.3.1](#)) aims to put this result into context with the literature. In 2008, by using functional site mapping studies, FH SCR-1/4 and SCR-19/20 were proposed to bind either the same C3b or two different C3b molecules ([Schmidt et al., 2008](#)). Following this, two structural studies concluded that they bind to either the same C3b molecule ([Morgan et al., 2011](#)) or two different C3b and C3d molecules ([Kajander et al., 2011](#)). The first study involved merging their C3d:SCR-19/20 crystal complex (PDB code: 3OXU) with an existing C3b:SCR-1/4 structure ([Wu et al., 2009](#)) and its concatenation to SCR-6/8 and SAXS-derived envelopes ([Morgan et](#)

[al., 2011](#)). The resulting full-length FH model was constructed by juxtaposition or superposition of the structures of SCR-5 and SCR-6/8 and the SAXS-shape envelopes of SCR-8/15 and SCR-15/19. This FH model was not subject to molecular dynamics or energy minimisation procedures. For the latter study, the crystal structure of SCR-19/20:C3d₂ showed that FH binding to two C3b molecules was sterically impossible, but that SCR-19/20 can simultaneously bind to C3b (the C3d part) and C3d ([Kajander et al., 2011](#)). The bidentate binding of C3b by SCR-19 (C3d part) and SCR-1/4 (via aNT, MG7, MG6, MG2, CUB, TED or C3d) ([Wu et al., 2009](#)) was proposed to enhance both the avidity and activity of FH, and SCR-20 was able to bind an earlier deposited C3d ([Kajander et al., 2011](#)). In 2017, another study of the crystal structure of the SCR-19/20, C3dg and OspE complex showed SCR-19 binding to C3dg (of either C3b or C3d) and SCR-20 binding to OspE. For SCR-20, in the absence of OspE, it may also bind another C3d ([Kajander et al., 2011](#)). Following these studies, both X-ray crystallography and SAXS were employed for the structural characterisation of SCR-18/20. The SAXS concluded that under solution conditions SCR-18/20 adopts multiple conformations mediated by the flexible linker between SCR-18 and SCR-19 ([Morgan et al., 2012](#)) (PDB code: 3SW0). In 2013, a mini-FH construct consisting of SCR-1/4 and SCR-19/20 bound together by a 12 glycine residue linker was engineered as a novel therapeutic agent ([Schmidt et al., 2013](#)). The structural model of the C3b:mini-FH complex was computationally constructed by superimposing the TED/C3d domain of the C3b:FH1-4 (PDB code: 2WII) ([Wu et al., 2009](#)) and C3d:FH19-20 (PDB code: 3OXU) ([Morgan et al., 2011](#)) complex structures via alignment of the common TED in C3b and C3d. The resulting model of the tertiary complex revealed that the two FH fragments were in a close proximity ([Schmidt et al., 2013](#)). In 2017, in another study, this mini-FH construct (SCR-1/4 and SCR-19/20) was co-crystallised with one C3b molecule and FI. The resulting mini-FH-FI-C3b complex revealed how FI generates the late-stage opsonins iC3b or C3dg by its cleavage activities. In this model of FH, SCR-19 binds to the TED of the shared C3b molecule ([Xue et al., 2017](#)) (PDB code: 5O32).

In summary, most crystal structures of FH in complex with C3b show both SCR-1/4 and SCR-19 binding to the same C3b molecule. Despite the high-resolution of X-ray crystallography for the structures of the FH fragments, the binding of glycosylated full-length FH to C3b had not been considered. Thus, the 1:1 C3b:FH model deduced from X-ray crystallography studies above may lack physiological relevance. In addition, the resulting full-length FH models do not use molecular dynamics or energy minimisation

procedures to either find physically plausible structures or statistically validate their modelling results. Furthermore, in the two most recent studies using the mini-FH construct, the ends of the modelled FH SCR-1/4 and SCR-19/20 fragments are artificially held together by 12 glycine residues.

In this PhD thesis, for both of the resulting full-length solution-state FH models, SCR-1/4 and SCR-19 are too far apart to bind the same C3b molecule ([Chapter 5, Figure 5.8](#)). By capturing full-length glycosylated FH in solution, the experimental conditions of the previous SAXS set up are more physiologically relevant. Thus, despite the lower resolution of SAXS when compared to X-ray crystallography, when combined with atomistic molecular modelling in this thesis, this allowed the flexibility of FH to be studied, and the molecular modelling identified an ensemble of physically plausible, energetically stable models that well fitted the SAXS data. However, further experimental studies are required to validate the outcome of the two major structural conformations of FH.

Overall, in order to verify the full-length structures of FH and the stoichiometry of the FH:C3b complex, further experiments using solution-based techniques such as SPR are required. For example, by using SPR, the binding of FH to C3b can be monitored, and the resulting data can be fitted to a binding model. In addition, the use of cryo-EM techniques may allow FH-C3b complexes, with full-length and glycosylated FH, to be observed directly in multiple conformations in their native environment at near atomic resolution. For cryo-EM, in order to reconstruct the 3D volume, complicated post-processing involving classifying, aligning and averaging by using statistics is required. However, compared to SAXS which is fast and easy but of lower resolution, a cryo-EM experiment yields a higher resolution 3D density map ([Kim et al., 2017](#)), which will help to verify the stoichiometry of the FH-C3b complex. These future projects will help to elucidate the impact of disease variants on the functioning of the FH-C3b complex, and facilitate the development of disease therapies that target the disease variants.

7.4 Structural analyses rationalise the distribution of aHUS and C3G rare variants in complement factor H

The aim of [Chapter 6](#) was to bring together the aHUS and C3G RV analyses from [Chapter 4](#) and the FH modelling from [Chapter 5](#) in order to identify whether the structural

location of a complement variant may predict the occurrence of aHUS or C3G in patients. For genetic variants that lead to changes in the protein, their locations in different types of functional sites will typically have different consequences. Thus, residues involved in molecular interactions are mainly surface-associated whereas residues involved in structural stabilisation are typically buried or surface-associated ([Chapter 2, Section 2.4](#)). The disruption of a structural disulphide bond often leads to LoF via protein misfolding. These LoF events can be predicted by the presence of rare missense variants in disulphide bond-forming Cys residues. The SCR domain is one of the most abundant domains in the complement proteins, occurring in FH, MCP and FB, and is structurally stabilised by two disulphide bridges formed by two pairs of conserved Cys residues ([Chapter 3, Section 3.3.8](#)). Previously, for the SCR consensus domain for FH and MCP, the hypervariable loop was suggested to have an association with AMD, aHUS and C3G disease variants ([Rodriguez et al., 2014](#)).

For [Chapter 6](#), the structural basis for the effect of predisposing variants for aHUS and C3G were reviewed by predicting their structural and functional impacts. For aHUS and C3G, the most common variants were non-truncating, missense changes in the regulators FH, FI and MCP, and the activators C3 and FB ([Chapter 4, Figure 4.3](#)). In [Chapter 6](#), their molecular correlations with aHUS and C3G were assessed by mapping the variants onto three-dimensional structures for the FH-C3b-FI, FH-C3dg and C3b-FB complexes and full-length FH. The 128 aHUS rare missense variants in FH mostly affected three types of residues, namely 13 case alleles in the 42 buried residues (31%), followed by 290 case alleles in the 1045 non-binding surface-accessible residues (28%), and finally 12 case alleles in the 92 residues at C3b-binding interfaces (13%), but none were found at the FI interfaces. In contrast, for 19 C3G variants in FH, these three types of residues totalled 2% (one case allele), 3% (26 case alleles) and 5% (five case alleles) respectively, and two case alleles (6%) within the FI binding interface were now involved. Another prominent group of variants involve Cys residues involved in disulphide bridges. Only those in the regulators FH, MCP and FI were associated with aHUS, while only those in FH and C3 were associated with C3G, when these were compared to the ExAC reference dataset. In conclusion, these differences between aHUS and C3G reflect their different pathologies, and importantly indicate that the structural location of a newly-discovered variant may predict the occurrence of aHUS or C3G in patients.

7.4.1 Future work

In future work, these residue type analyses could be extended to the other complement proteins that have high frequencies of rare missense variants for aHUS and C3G. These are C3, FI, MCP and FB, in their respective regulatory or activating complement complexes. In order to correlate rare missense variants that predispose for aHUS and C3G with different structural functions, their effects on protein stability could be analysed in more detail by using free energy change calculations and statistics. These analyses could also be extended to new rare missense variant data for AMD patients. Following comparisons with aHUS and C3G, novel insights into the differences between the molecular bases of the three inflammatory diseases may be identified.

- Abrera-Abeleda, M. A., Nishimura, C., Frees, K., Jones, M., Maga, T., Katz, L. M., Zhang, Y. and Smith, R. J. (2011) Allelic variants of complement genes associated with dense deposit disease. *J. Am. Soc. Nephrol.* **22**, 1551-1559
- Adams, T. E., Li, W. and Huntington, J. A. (2009) Molecular basis of thrombomodulin activation of slow thrombin. *J. Thromb. Haemost.* **7**, 1688-1695
- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. and Sunyaev, S. R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods.* **7**, 248-249
- Akira, S., Uematsu, S. and Takeuchi, O. (2006) Pathogen recognition and innate immunity. *Cell.* **124**, 783-801
- Alberts, B. (2002) *Molecular biology of the cell*, 4th Ed., Garland Science, New York
- Alder, B. J. and Wainwright, T. E. (1959) Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.* **31**, 459-466
- Alicic, R. Z., Rooney, M. T. and Tuttle, K. R. (2017) Diabetic Kidney Disease: Challenges, Progress, and Possibilities. *Clin. J. Am. Soc. Nephrol.* **12**, 2032-2045
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410
- Altshuler, D., Daly, M. J. and Lander, E. S. (2008) Genetic mapping in human disease. *Science.* **322**, 881-888
- Amaral, M. M., Sacerdoti, F., Jancic, C., Repetto, H. A., Paton, A. W., Paton, J. C. and Ibarra, C. (2013) Action of shiga toxin type-2 and subtilase cytotoxin on human microvascular endothelial cells. *PLoS One.* **8**, e70431
- Anderson, D. H., Radeke, M. J., Gallo, N. B., Chapin, E. A., Johnson, P. T., Curletti, C. R., Hancox, L. S., Hu, J., Ebright, J. N., Malek, G., Hauser, M. A., Rickman, C. B., Bok, D., Hageman, G. S. and Johnson, L. V. (2010) The pivotal role of the complement system in aging and age-related macular degeneration: hypothesis re-visited. *Prog. Retin. Eye Res.* **29**, 95-112
- Anfinsen, C. B. (1973) Principles that govern the folding of protein chains. *Science.* **181**, 223-230
- Apic, G., Gough, J. and Teichmann, S. A. (2001) An insight into domain combinations. *Bioinformatics.* **17 Suppl 1**, S83-89
- Appel, G. B., Cook, H. T., Hageman, G., Jennette, J. C., Kashgarian, M., Kirschfink, M., Lambris, J. D., Lanning, L., Lutz, H. U., Meri, S., Rose, N. R., Salant, D. J., Sethi, S., Smith, R. J., Smoyer, W., Tully, H. F., Tully, S. P., Walker, P., Welsh, M., Wurzner, R. and Zipfel, P. F. (2005) Membranoproliferative glomerulonephritis type II (dense deposit disease): an update. *J. Am. Soc. Nephrol.* **16**, 1392-1403
- Apte, R. S. (2016) Targeting Tissue Lipids in Age-related Macular Degeneration. *EBioMedicine.* **5**, 26-27
- Armstrong, C. T., Mason, P. E., Anderson, J. L. and Dempsey, C. E. (2016) Arginine side chain interactions and the role of arginine as a gating charge carrier in voltage sensitive ion channels. *Sci. Rep.* **6**, 21759
- Armstrong, R. A. (2014) When to use the Bonferroni correction. *Ophthalmic Physiol. Opt.* **34**, 502-508
- Arsenic, R., Treue, D., Lehmann, A., Hummel, M., Dietel, M., Denkert, C. and Budczies, J. (2015) Comparison of targeted next-generation sequencing and Sanger sequencing for the detection of PIK3CA mutations in breast cancer. *BMC Clin. Pathol.* **15**, 20
- Aslam, M. and Perkins, S. J. (2001) Folded-back solution structure of monomeric factor H of human complement by synchrotron X-ray and neutron scattering, analytical

- ultracentrifugation and constrained molecular modelling. *J. Mol. Biol.* **309**, 1117-1138
- Athanasiou, Y., Voskarides, K., Gale, D. P., Damianou, L., Patsias, C., Zavros, M., Maxwell, P. H., Cook, H. T., Demosthenous, P., Hadjisavvas, A., Kyriacou, K., Zouvani, I., Pierides, A. and Deltas, C. (2011) Familial C3 glomerulopathy associated with CFHR5 mutations: clinical characteristics of 91 patients in 16 pedigrees. *Clin. J. Am. Soc. Nephrol.* **6**, 1436-1446
- Auer, P. L., Johnsen, J. M., Johnson, A. D., Logsdon, B. A., Lange, L. A., Nalls, M. A., Zhang, G., Franceschini, N., Fox, K., Lange, E. M., Rich, S. S., O'Donnell, C. J., Jackson, R. D., Wallace, R. B., Chen, Z., Graubert, T. A., Wilson, J. G., Tang, H., Lettre, G., Reiner, A. P., Ganesh, S. K. and Li, Y. (2012) Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *Am. J. Hum. Genet.* **91**, 794-808
- Auer, P. L. and Lettre, G. (2015) Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7**, 16
- Ault, B. H., Schmidt, B. Z., Fowler, N. L., Kashtan, C. E., Ahmed, A. E., Vogt, B. A. and Colten, H. R. (1997) Human factor H deficiency. Mutations in framework cysteine residues and block in H protein secretion and intracellular catabolism. *J. Biol. Chem.* **272**, 25168-25175
- Azzi, A., Boscoboinik, D. and Hensey, C. (1992) The protein kinase C family. *Eur. J. Biochem.* **208**, 547-557
- Babanejad, M., Moein, H., Akbari, M. R., Badiei, A., Yaseri, M., Soheilian, M. and Najmabadi, H. (2016) Investigating the CFH Gene Polymorphisms as a Risk Factor for Age-related Macular Degeneration in an Iranian Population. *Ophthalmic Genet.* **37**, 144-149
- Bacaër, N. (2011) *The Hardy–Weinberg law (1908)*, Springer Verlag, London
- Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. and Yeh, L. S. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154-159
- Bajwa, R., DePalma, J. A., Khan, T., Cheema, A., Kalathil, S. A., Hossain, M. A., Haroon, A., Madhurima, A., Zheng, M., Nayer, A. and Asif, A. (2018) C3 Glomerulopathy and Atypical Hemolytic Uremic Syndrome: Two Important Manifestations of Complement System Dysfunction. *Case Rep. Nephrol. Dial.* **8**, 25-34
- Baker, E. N. and Hubbard, R. E. (1984) Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97-179
- Barbour, T. D., Pickering, M. C. and Terence Cook, H. (2013) Dense deposit disease and C3 glomerulopathy. *Semin. Nephrol.* **33**, 493-507
- Barc, J. and Koopmann, T. T. (2011) Genome-wide association studies: providers of candidate genes for identification of rare variants? *EP Eur.* **13**, 911-912
- Barlow, P. N., Norman, D. G., Steinkasserer, A., Horne, T. J., Pearce, J., Driscoll, P. C., Sim, R. B. and Campbell, I. D. (1992) Solution structure of the fifth repeat of factor H: a second example of the complement control protein module. *Biochemistry.* **31**, 3626-3634
- Barlow, P. N., Steinkasserer, A., Norman, D. G., Kieffer, B., Wiles, A. P., Sim, R. B. and Campbell, I. D. (1993) Solution structure of a pair of complement modules by nuclear magnetic resonance. *J. Mol. Biol.* **232**, 268-284
- Barrett, J. C., et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955-962

- Bartlett, A. I. and Radford, S. E. (2009) An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. *Nat. Struct. Mol. Biol.* **16**, 582-588
- Barton, N. H. (2010) Mutation and the evolution of recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 1281-1294
- Beamish, C. (2001) *Transversion Mutation A2 - Brenner, Sydney.* in *Encyclopedia of Genetics* (Miller, J. H. ed.), Academic Press, New York. pp 2042
- Beaudet, A. L. and Tsui, L. C. (1993) A suggested nomenclature for designating mutations. *Hum. Mutat.* **2**, 245-248
- Bennett, C. A., Petrovski, S., Oliver, K. L. and Berkovic, S. F. (2017) ExACtly zero or once: A clinically helpful guide to assessing genetic variants in mild epilepsies. *Neurol. Genet.* **3**, e163
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2008) GenBank. *Nucleic Acids Res.* **36**, D25-30
- Benson, N. C. and Daggett, V. (2012) A comparison of multiscale methods for the analysis of molecular dynamics simulations. *J. Phys. Chem. B.* **116**, 8722-8731
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242
- Bernado, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. and Svergun, D. I. (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* **129**, 5656-5664
- Beutler, B. (2004) Innate immunity: an overview. *Mol. Immunol.* **40**, 845-859
- Beutler, E. (1993) The designation of mutations. *Am. J. Hum. Genet.* **53**, 783-785
- Bexborn, F., Andersson, P. O., Chen, H., Nilsson, B. and Ekdahl, K. N. (2008) The tick-over theory revisited: formation and regulation of the soluble alternative complement C3 convertase (C3(H₂O)Bb). *Mol. Immunol.* **45**, 2370-2379
- Bhattacharjee, A., Reuter, S., Trojnar, E., Kolodziejczyk, R., Seeberger, H., Hyvarinen, S., Uzonyi, B., Szilagyi, A., Prohaszka, Z., Goldman, A., Jozsi, M. and Jokiranta, T. S. (2015) The major autoantibody epitope on factor H in atypical hemolytic uremic syndrome is structurally different from its homologous site in factor H-related protein 1, supporting a novel model for induction of autoimmunity in this disease. *J. Biol. Chem.* **290**, 9500-9510
- Biedermannova, L. and Schneider, B. (2015) Structure of the ordered hydration of amino acids in proteins: analysis of crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **71**, 2192-2202
- Bird, A. C. (2010) Therapeutic targets in age-related macular disease. *J. Clin. Invest.* **120**, 3033-3041
- Black, J. R. and Clark, S. J. (2016) Age-related macular degeneration: genome-wide association studies to translation. *Genet. Med.* **18**, 283-289
- Blackmore, T. K., Hellwage, J., Sadlon, T. A., Higgs, N., Zipfel, P. F., Ward, H. M. and Gordon, D. L. (1998) Identification of the second heparin-binding domain in human complement factor H. *J. Immunol.* **160**, 3342-3348
- Blackmore, T. K., Sadlon, T. A., Ward, H. M., Lublin, D. M. and Gordon, D. L. (1996) Identification of a heparin binding domain in the seventh short consensus repeat of complement factor H. *J. Immunol.* **157**, 5422-5427
- Blaum, B. S., Hannan, J. P., Herbert, A. P., Kavanagh, D., Uhrin, D. and Stehle, T. (2015) Structural basis for sialic acid-mediated self-recognition by complement factor H. *Nat. Chem. Biol.* **11**, 77-82
- Blom, A. M., Kask, L. and Dahlback, B. (2001) Structural requirements for the complement regulatory activities of C4BP. *J. Biol. Chem.* **276**, 27136-27144

- Blom, A. M., Kask, L. and Dahlback, B. (2003) CCP1-4 of the C4b-binding protein alpha-chain are required for factor I mediated cleavage of complement factor C3b. *Mol. Immunol.* **39**, 547-556
- Blom, A. M., Villoutreix, B. O. and Dahlback, B. (2004) Complement inhibitor C4b-binding protein-friend or foe in the innate immune system? *Mol. Immunol.* **40**, 1333-1346
- Blom, A. M., Webb, J., Villoutreix, B. O. and Dahlback, B. (1999) A cluster of positively charged amino acids in the C4BP alpha-chain is crucial for C4b binding and factor I cofactor function. *J. Biol. Chem.* **274**, 19237-19245
- Bok, D. (2005) Evidence for an inflammatory process in age-related macular degeneration gains new support. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7053-7054
- Bokisch, V. A., Dierich, M. P. and Muller-Eberhard, H. J. (1975) Third component of complement (C3): structural properties in relation to functions. *Proc. Natl. Acad. Sci. U.S.A.* **72**, 1989-1993
- Boycott, K. M., Vanstone, M. R., Bulman, D. E. and MacKenzie, A. E. (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* **14**, 681-691
- Branden, C. and Tooze, J. (1999) *Introduction to protein structure*, 2nd ed. Ed., Garland Pub, New York
- Breda, L. C., Hsieh, C. L., Castiblanco Valencia, M. M., da Silva, L. B., Barbosa, A. S., Blom, A. M., Chang, Y. F. and Isaac, L. (2015) Fine Mapping of the Interaction between C4b-Binding Protein and Outer Membrane Proteins LigA and LigB of Pathogenic *Leptospira interrogans*. *PLoS Negl. Trop. Dis.* **9**, e0004192
- Bresin, E., Rurali, E., Caprioli, J., Sanchez-Corral, P., Fremeaux-Bacchi, V., Rodriguez de Cordoba, S., Pinto, S., Goodship, T. H., Alberti, M., Ribes, D., Valoti, E., Remuzzi, G., Noris, M. and European Working Party on Complement Genetics in Renal, D. (2013) Combined complement gene mutations in atypical hemolytic uremic syndrome influence clinical phenotype. *J. Am. Soc. Nephrol.* **24**, 475-486
- Brookes, A. J. (1999) The essence of SNPs. *Gene.* **234**, 177-186
- Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M. and Karplus, M. (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30**, 1545-1614
- Bruel, A., Kavanagh, D., Noris, M., Delmas, Y., Wong, E. K. S., Bresin, E., Provot, F., Brocklebank, V., Mele, C., Remuzzi, G., Loirat, C., Fremeaux-Bacchi, V. and Fakhouri, F. (2017) Hemolytic uremic syndrome in pregnancy and postpartum. *Clin. J. Am. Soc. Nephrol.*
- Bu, F., Borsa, N., Gianluigi, A. and Smith, R. J. (2012) Familial atypical hemolytic uremic syndrome: a review of its genetic and clinical aspects. *Clin. Dev. Immunol.* **2012**, 370426
- Bu, F., Borsa, N. G., Jones, M. B., Takanami, E., Nishimura, C., Hauer, J. J., Azaiez, H., Black-Ziegelbein, E. A., Meyer, N. C., Kolbe, D. L., Li, Y., Frees, K., Schnieders, M. J., Thomas, C., Nester, C. and Smith, R. J. (2016) High-throughput genetic testing for thrombotic microangiopathies and C3 glomerulopathies. *J. Am. Soc. Nephrol.* **27**, 1245-1253
- Bu, F., Maga, T., Meyer, N. C., Wang, K., Thomas, C. P., Nester, C. M. and Smith, R. J. (2014) Comprehensive genetic analysis of complement and coagulation genes in atypical hemolytic uremic syndrome. *J. Am. Soc. Nephrol.* **25**, 55-64

- Buchner, H. (1891) Zur Nomenklatur der schützenden Eiweisskörper. *Centr Bakteriol Parasitenk.* 699-701
- Bush, W. S. and Moore, J. H. (2012) Chapter 11: Genome-Wide Association Studies. *PLoS Comp. Biol.* **8**, e1002822
- Campbell, R. D., Dodds, A. W. and Porter, R. R. (1980) The binding of human complement component C4 to antibody-antigen aggregates. *Biochem. J.* **189**, 67-80
- Caprioli, J., Noris, M., Brioschi, S., Pianetti, G., Castelletti, F., Bettinaglio, P., Mele, C., Bresin, E., Cassis, L., Gamba, S., Porrati, F., Bucchioni, S., Monteferrante, G., Fang, C. J., Liszewski, M. K., Kavanagh, D., Atkinson, J. P., Remuzzi, G., International Registry of, R. and Familial, H. T. (2006) Genetics of HUS: the impact of MCP, CFH, and IF mutations on clinical presentation, response to treatment, and outcome. *Blood.* **108**, 1267-1279
- Carroni, M. and Saibil, H. R. (2016) Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods.* **95**, 78-85
- Case, D. A., Cheatham, T. E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr., Onufriev, A., Simmerling, C., Wang, B. and Woods, R. J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668-1688
- Chamberlain, D., Ullman, C. G. and Perkins, S. J. (1998) Possible arrangement of the five domains in human complement factor I as determined by a combination of X-ray and neutron scattering and homology modeling. *Biochemistry.* **37**, 13918-13929
- Chaplin, D. D. (2010) Overview of the immune response. *J. Allergy Clin. Immunol.* **125**, S3-23
- Chaudhary, P., Hepgur, M., Sarkissian, S., Smith, R. J. and Weitz, I. C. (2014) Atypical haemolytic-uraemic syndrome due to heterozygous mutations of CFH/CFHR1-3 and complement factor H 479. *Blood Transfus.* **12**, 111-113
- Chauvet, S., Roumenina, L. T., Bruneau, S., Marinozzi, M. C., Rybkine, T., Schramm, E. C., Java, A., Atkinson, J. P., Aldigier, J. C., Bridoux, F., Touchard, G. and Fremeaux-Bacchi, V. (2016) A familial C3GN secondary to defective C3 regulation by complement receptor 1 and complement factor H. *J. Am. Soc. Nephrol.* **27**, 1665-1677
- Chen, Q., Wiesener, M., Eberhardt, H. U., Hartmann, A., Uzonyi, B., Kirschfink, M., Amann, K., Buettner, M., Goodship, T., Hugo, C., Skerka, C. and Zipfel, P. F. (2014) Complement factor H-related hybrid protein deregulates complement in dense deposit disease. *J. Clin. Invest.* **124**, 145-155
- Chen, R., Shi, L., Hakenberg, J., Naughton, B., Sklar, P., Zhang, J., Zhou, H., Tian, L., Prakash, O., Lemire, M., Sleiman, P., Cheng, W. Y., Chen, W., Shah, H., Shen, Y., Fromer, M., Omberg, L., Deardorff, M. A., Zackai, E., Bobe, J. R., Levin, E., Hudson, T. J., Groop, L., Wang, J., Hakonarson, H., Wojcicki, A., Diaz, G. A., Edelman, L., Schadt, E. E. and Friend, S. H. (2016) Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat. Biotechnol.* **34**, 531-538
- Choi, E. J. and Mayo, S. L. (2006) Generation and analysis of proline mutants in protein G. *Protein Eng. Des. Sel.* **19**, 285-289
- Chothia, C. and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826
- Claesen, J. and Burzykowski, T. (2017) Computational methods and challenges in hydrogen/deuterium exchange mass spectrometry. *Mass Spectrom. Rev.* **36**, 649-667

- Clark, R. A. and Klebanoff, S. J. (1978) Role of the classical and alternative complement pathways in chemotaxis and opsonization: studies of human serum deficient in C4. *J. Immunol.* **120**, 1102-1108
- Clark, S. J. and Bishop, P. N. (2015) Role of Factor H and Related Proteins in Regulating Complement Activation in the Macula, and Relevance to Age-Related Macular Degeneration. *J. Clin. Med.* **4**, 18-31
- Clark, S. J., Perveen, R., Hakobyan, S., Morgan, B. P., Sim, R. B., Bishop, P. N. and Day, A. J. (2010) Impaired binding of the age-related macular degeneration-associated complement factor H 402H allotype to Bruch's membrane in human retina. *J. Biol. Chem.* **285**, 30192-30202
- Clark, S. J., Ridge, L. A., Herbert, A. P., Hakobyan, S., Mulloy, B., Lennon, R., Wurznier, R., Morgan, B. P., Uhrin, D., Bishop, P. N. and Day, A. J. (2013) Tissue-specific host recognition by complement factor H is mediated by differential activities of its glycosaminoglycan-binding regions. *J. Immunol.* **190**, 2049-2057
- Clark, S. J., Schmidt, C. Q., White, A. M., Hakobyan, S., Morgan, B. P. and Bishop, P. N. (2014) Identification of factor H-like protein 1 as the predominant complement regulator in Bruch's membrane: implications for age-related macular degeneration. *J. Immunol.* **193**, 4962-4970
- Columb, M. O. and Atkinson, M. S. (2016) Statistical analysis: sample size and power estimations. *BJA Education.* **16**, 159-161
- Conway, E. M. (2012) Thrombomodulin and its role in inflammation. *Semin. Immunopathol.* **34**, 107-125
- Cook, H. T. (2017) C3 glomerulopathy. *FI000Res.* **6**, 248
- Crabb, J. W., Miyagi, M., Gu, X., Shadrach, K., West, K. A., Sakaguchi, H., Kamei, M., Hasan, A., Yan, L., Rayborn, M. E., Salomon, R. G. and Hollyfield, J. G. (2002) Drusen proteome analysis: an approach to the etiology of age-related macular degeneration. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 14682-14687
- Crick, F. H. (1966) Codon--anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* **19**, 548-555
- Csincsi, A. I., Kopp, A., Zoldi, M., Banlaki, Z., Uzonyi, B., Hebecker, M., Caesar, J. J., Pickering, M. C., Daigo, K., Hamakubo, T., Lea, S. M., Goicoechea de Jorge, E. and Jozsi, M. (2015) Factor H-related protein 5 interacts with pentraxin 3 and the extracellular matrix and modulates complement activation. *J. Immunol.* **194**, 4963-4973
- Curtis, J. E., Raghunandan, S., Nanda, H. and Krueger, S. (2012) SASSIE: A program to study intrinsically disordered biological molecules and macromolecular ensembles using experimental scattering restraints. *Comput. Phys. Commun.* **183**, 382-389
- Darwin, C. (1859) *On the Origin of Species by means of Natural Selection, or the preservation of favoured races in the struggle for life*, John Murray, London
- Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M. D., Bhat, D., Chivian, D., Kim, D. E., Sheffler, W. H., Malmstrom, L., Wollacott, A. M., Wang, C., Andre, I. and Baker, D. (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins.* **69 Suppl 8**, 118-128
- David, C. C. and Jacobs, D. J. (2014) Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol. Biol.* **1084**, 193-226
- Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A. and Sillitoe, I. (2017) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **45**, D289-D295

- Delvaeye, M., Noris, M., De Vriese, A., Esmon, C. T., Esmon, N. L., Ferrell, G., Del-Favero, J., Plaisance, S., Claes, B., Lambrechts, D., Zoja, C., Remuzzi, G. and Conway, E. M. (2009) Thrombomodulin mutations in atypical hemolytic-uremic syndrome. *N. Engl. J. Med.* **361**, 345-357
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernysky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. and Daly, M. J. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498
- Dillon, O. J., Lunke, S., Stark, Z., Yeung, A., Thorne, N., Melbourne Genomics Health, A., Gaff, C., White, S. M. and Tan, T. Y. (2018) Exome sequencing has higher diagnostic yield compared to simulated disease-specific panels in children with suspected monogenic disorders. *Eur. J. Hum. Genet.* **26**, 644-651
- DiScipio, R. G. (1992) Ultrastructures and interactions of complement factors H and I. *J. Immunol.* **149**, 2592-2599
- Dodge, J. A., Chigladze, T., Donadieu, J., Grossman, Z., Ramos, F., Serlicorni, A., Siderius, L., Stefanidis, C. J., Tasic, V., Valiulis, A. and Wierzba, J. (2011) The importance of rare diseases: from the gene to society. *Arch. Dis. Child.* **96**, 791-792
- Dragon-Durey, M. A., Fremeaux-Bacchi, V., Loirat, C., Blouin, J., Niaudet, P., Deschenes, G., Coppo, P., Herman Fridman, W. and Weiss, L. (2004) Heterozygous and homozygous factor h deficiencies associated with hemolytic uremic syndrome or membranoproliferative glomerulonephritis: report and genetic analysis of 16 cases. *J. Am. Soc. Nephrol.* **15**, 787-795
- Dunne, O. (2015) Functional interactions of the C-terminus of the complement regulator Factor H with its ligands. *PhD Thesis, University College London.*
- Dupont, W. D. and Plummer, W. D., Jr. (1990) Power and sample size calculations. A review and computer program. *Control. Clin. Trials.* **11**, 116-128
- Durey, M. A., Sinha, A., Togarsimalemath, S. K. and Bagga, A. (2016) Anti-complement-factor H-associated glomerulopathies. *Nat. Rev. Nephrol.* **12**, 563-578
- Eberhardt, H. E., Chen, Q., Zipfel, P. and Skerka, C. (2011) Human complement factor H-related protein 2 (CFHR2) represents a novel complement regulator, which is reduced in a patient with MPGN I. *Mol. Immunol.* **48**, 1674-1674
- Eberhardt, H. U., Skerka, C., Zipfel, P. F., Hallstrom, T., Hartmann, A. and Chen, Q. (2012) C3-glomerulopathy associated human factor H-related proteins 2 (CFHR2) and 5 (CFHR5) regulate complement C3b and TCC. *Immunobiology.* **217**, 1143-1143
- Edwards, A. O., Ritter, R., Abel, K. J., Manning, A., Panhuysen, C. and Farrer, L. A. (2005) Complement factor H polymorphism and age-related macular degeneration. *Science.* **308**, 421-424
- Elber, R. and Karplus, M. (1987) Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science.* **235**, 318-321
- Elvington, M., Liszewski, M. K. and Atkinson, J. P. (2016) Evolution of the complement system: from defense of the single cell to guardian of the intravascular space. *Immunol. Rev.* **274**, 9-15
- Ermert, D., Weckel, A., Agarwal, V., Frick, I. M., Bjorck, L. and Blom, A. M. (2013) Binding of complement inhibitor C4b-binding protein to a highly virulent *Streptococcus pyogenes* M1 strain is mediated by protein H and enhances adhesion to and invasion of endothelial cells. *J. Biol. Chem.* **288**, 32172-32183

- Ermini, L., Goodship, T. H., Strain, L., Weale, M. E., Sacks, S. H., Cordell, H. J., Fremeaux-Bacchi, V. and Sheerin, N. S. (2012) Common genetic variants in complement genes other than CFH, CD46 and the CFHRs are not associated with aHUS. *Mol. Immunol.* **49**, 640-648
- Ernst, C., Hahnen, E., Engel, C., Nothnagel, M., Weber, J., Schmutzler, R. K. and Hauke, J. (2018) Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. *BMC Med. Genomics.* **11**, 35
- Ernst, J. A., Clubb, R. T., Zhou, H. X., Gronenborn, A. M. and Clore, G. M. (1995) Demonstration of positionally disordered water within a protein hydrophobic cavity by NMR. *Science.* **267**, 1813-1817
- Esparza-Gordillo, J., Goicoechea de Jorge, E., Buil, A., Carreras Berges, L., Lopez-Trascasa, M., Sanchez-Corral, P. and Rodriguez de Cordoba, S. (2005) Predisposition to atypical hemolytic uremic syndrome involves the concurrence of different susceptibility alleles in the regulators of complement activation gene cluster in 1q32. *Hum. Mol. Genet.* **14**, 703-712
- Esparza-Gordillo, J., Goicoechea de Jorge, E., Garrido, C. A., Carreras, L., Lopez-Trascasa, M., Sanchez-Corral, P. and Rodriguez de Cordoba, S. (2006) Insights into hemolytic uremic syndrome: segregation of three independent predisposition factors in a large, multiple affected pedigree. *Mol. Immunol.* **43**, 1769-1775
- Evangelou, E., Trikalinos, T. A., Salanti, G. and Ioannidis, J. P. (2006) Family-based versus unrelated case-control designs for genetic associations. *PLoS Genet.* **2**, e123
- Eyler, S. J., Meyer, N. C., Zhang, Y., Xiao, X., Nester, C. M. and Smith, R. J. (2013) A novel hybrid CFHR1/CFH gene causes atypical hemolytic uremic syndrome. *Pediatr. Nephrol.* **28**, 2221-2225
- Fakhouri, F., Fremeaux-Bacchi, V., Noel, L. H., Cook, H. T. and Pickering, M. C. (2010) C3 glomerulopathy: a new classification. *Nat. Rev. Nephrol.* **6**, 494-499
- Fakhouri, F., Zuber, J., Fremeaux-Bacchi, V. and Loirat, C. (2017) Haemolytic uraemic syndrome. *Lancet.*
- Farries, T. C., Seya, T., Harrison, R. A. and Atkinson, J. P. (1990) Competition for binding sites on C3b by CR1, CR2, MCP, factor B and factor H. *Complement Inflamm.* **7**, 30-41
- Fearon, D. T. (1977) Purification of C3b inactivator and demonstration of its two polypeptide chain structure. *J. Immunol.* **119**, 1248-1252
- Fearon, D. T. and Austen, K. F. (1975) Properdin: binding to C3b and stabilization of the C3b-dependent C3 convertase. *J. Exp. Med.* **142**, 856-863
- Fenaille, F., Le Mignon, M., Groseil, C., Ramon, C., Riande, S., Siret, L. and Bihoreau, N. (2007) Site-specific N-glycan characterization of human complement factor H. *Glycobiology.* **17**, 932-944
- Fernando, A. N., Furtado, P. B., Clark, S. J., Gilbert, H. E., Day, A. J., Sim, R. B. and Perkins, S. J. (2007) Associative and structural properties of the region of complement factor H encompassing the Tyr402His disease-related polymorphism and its interactions with heparin. *J. Mol. Biol.* **368**, 564-581
- Ferrara, P., Apostolakis, J. and Caflisch, A. (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins.* **46**, 24-33
- Ferreira, V. P., Pangburn, M. K. and Cortes, C. (2010) Complement control protein factor H: the good, the bad, and the inadequate. *Mol. Immunol.* **47**, 2187-2197
- Finkelstein, A. V. (2018) 50+ Years of Protein Folding. *Biochemistry (Mosc.).* **83**, S3-S18

- Fishelson, Z., Pangburn, M. K. and Muller-Eberhard, H. J. (1984) Characterization of the initial C3 convertase of the alternative pathway of human complement. *J. Immunol.* **132**, 1430-1434
- Fisher, R. A. S. (1925) *Statistical methods for Research Workers*, Oliver and Boyd, Edinburgh
- Flajnik, M. F. and Kasahara, M. (2010) Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* **11**, 47-59
- Flanagan, S. E., Patch, A.-M. and Ellard, S. (2010) Using SIFT and PolyPhen to Predict Loss-of-Function and Gain-of-Function Mutations. *Genet. Test. Mol. Biomarkers.* **14**, 533-537
- Flicek, P., et al. (2008) Ensembl 2008. *Nucleic Acids Res.* **36**, D707-714
- Forneris, F., Ricklin, D., Wu, J., Tzekou, A., Wallace, R. S., Lambris, J. D. and Gros, P. (2010) Structures of C3b in complex with factors B and D give insight into complement convertase formation. *Science.* **330**, 1816-1820
- Forneris, F., Wu, J., Xue, X., Ricklin, D., Lin, Z., Sfyroera, G., Tzekou, A., Volokhina, E., Granneman, J. C., Hauhart, R., Bertram, P., Liszewski, M. K., Atkinson, J. P., Lambris, J. D. and Gros, P. (2016) Regulators of complement activity mediate inhibitory mechanisms through a common C3b-binding mode. *EMBO J.* **35**, 1133-1149
- Frank, M. M. and Sullivan, K. E. (2014) *Chapter 38 - Deficiencies of the Complement System.* in *Stiehm's Immune Deficiencies*, Academic Press, Amsterdam. pp 731-763
- Frauenfelder, H., Sligar, S. G. and Wolynes, P. G. (1991) The energy landscapes and motions of proteins. *Science.* **254**, 1598-1603
- Frazer, K. A., Sheehan, J. B., Stokowski, R. P., Chen, X., Hosseini, R., Cheng, J. F., Fodor, S. P., Cox, D. R. and Patil, N. (2001) Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* **11**, 1651-1659
- Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter, N. P., Scherer, S. W. and Lee, C. (2006) Copy number variation: new insights in genome diversity. *Genome Res.* **16**, 949-961
- Fremaux-Bacchi, V., Fakhouri, F., Garnier, A., Benaïme, F., Dragon-Durey, M. A., Ngo, S., Moulin, B., Servais, A., Provot, F., Rostaing, L., Burtey, S., Niaudet, P., Deschenes, G., Lebranchu, Y., Zuber, J. and Loirat, C. (2013) Genetics and outcome of atypical hemolytic uremic syndrome: a nationwide French series comparing children and adults. *Clin. J. Am. Soc. Nephrol.* **8**, 554-562
- Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., Project, N. E. S. and Akey, J. M. (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature.* **493**, 216-220
- Fuentes-Prior, P., Iwanaga, Y., Huber, R., Pagila, R., Rumennik, G., Seto, M., Morser, J., Light, D. R. and Bode, W. (2000) Structural basis for the anticoagulant activity of the thrombin-thrombomodulin complex. *Nature.* **404**, 518-525
- Fujita, T. and Nussenzweig, V. (1979) The role of C4-binding protein and beta 1H in proteolysis of C4b and C3b. *J. Exp. Med.* **150**, 267-276
- Fung, K. W., Wright, D. W., Gor, J., Swann, M. J. and Perkins, S. J. (2016) Domain structure of human complement C4b extends with increasing NaCl concentration: implications for its regulatory mechanism. *Biochem. J.* **473**, 4473-4491
- Gale, D. P., de Jorge, E. G., Cook, H. T., Martinez-Barricarte, R., Hadjisavvas, A., McLean, A. G., Pusey, C. D., Pierides, A., Kyriacou, K., Athanasiou, Y., Voskarides, K., Deltas, C., Palmer, A., Fremaux-Bacchi, V., de Cordoba, S. R.,

- Maxwell, P. H. and Pickering, M. C. (2010) Identification of a mutation in complement factor H-related protein 5 in patients of Cypriot origin with glomerulonephritis. *Lancet*. **376**, 794-801
- Ganu, V. S., Muller-Eberhard, H. J. and Hugli, T. E. (1989) Factor C3f is a spasmogenic fragment released from C3b by factors I and H: the heptadecapeptide C3f was synthesized and characterized. *Mol. Immunol.* **26**, 939-948
- Garcia-de la Torre, J., Huertas, M. L. and Carrasco, B. (2000) Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophys. J.* **78**, 719-730
- Garred, P., Genster, N., Pilely, K., Bayarri-Olmos, R., Rosbjerg, A., Ma, Y. J. and Skjoedt, M. O. (2016) A journey through the lectin pathway of complement-MBL and beyond. *Immunol. Rev.* **274**, 74-97
- Geerlings, M. J., de Jong, E. K. and den Hollander, A. I. (2017) The complement system in age-related macular degeneration: A review of rare genetic variants and implications for personalized treatment. *Mol. Immunol.* **84**, 65-76
- Geerlings, M. J., Volokhina, E. B., de Jong, E. K., van de Kar, N., Pauper, M., Hoyng, C. B., van den Heuvel, L. P. and den Hollander, A. I. (2018) Genotype-phenotype correlations of low-frequency variants in the complement system in renal disease and age-related macular degeneration. *Clin. Genet.*
- Gemmell, N. J. and Slate, J. (2006) Heterozygote advantage for fecundity. *PLoS One*. **1**, e125
- Gendoo, D. M. and Harrison, P. M. (2012) The landscape of the prion protein's structural response to mutation revealed by principal component analysis of multiple NMR ensembles. *PLoS Comput. Biol.* **8**, e1002646
- Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. and McVean, G. A. (2010) A map of human genome variation from population-scale sequencing. *Nature*. **467**, 1061-1073
- Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A. and Abecasis, G. R. (2015) A global reference for human genetic variation. *Nature*. **526**, 68-74
- Gibson, G. (2011) Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135-145
- Gilissen, C., Hoischen, A., Brunner, H. G. and Veltman, J. A. (2011) Unlocking Mendelian disease using exome sequencing. *Genome Biol.* **12**, 228
- Glatter, O. (1977) A new method for the evaluation of small-angle scattering data. *J. Appl. Crystallogr.* **10**, 415-421
- Go, N. and Abe, H. (1981) Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers*. **20**, 991-1011
- Goicoechea de Jorge, E., Caesar, J. J., Malik, T. H., Patel, M., Colledge, M., Johnson, S., Hakobyan, S., Morgan, B. P., Harris, C. L., Pickering, M. C. and Lea, S. M. (2013) Dimerization of complement factor H-related proteins modulates complement activation in vivo. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 4685-4690
- Goicoechea de Jorge, E., Harris, C. L., Esparza-Gordillo, J., Carreras, L., Arranz, E. A., Garrido, C. A., Lopez-Trascasa, M., Sanchez-Corral, P., Morgan, B. P. and Rodriguez de Cordoba, S. (2007) Gain-of-function mutations in complement factor B are associated with atypical hemolytic uremic syndrome. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 240-245
- Goicoechea de Jorge, E. and Pickering, M. C. (2010) Atypical hemolytic uremic syndrome: telling the difference between H and Y. *Kidney Int.* **78**, 721-723

- Goldberger, G., Arnaout, M. A., Aden, D., Kay, R., Rits, M. and Colten, H. R. (1984) Biosynthesis and postsynthetic processing of human C3b/C4b inactivator (factor I) in three hepatoma cell lines. *J. Biol. Chem.* **259**, 6492-6497
- Goldenberg, D. P. and Argyle, B. (2014) Self crowding of globular proteins studied by small-angle x-ray scattering. *Biophys. J.* **106**, 895-904
- Goodship, T. H., Cook, H. T., Fakhouri, F., Fervenza, F. C., Fremeaux-Bacchi, V., Kavanagh, D., Nester, C. M., Noris, M., Pickering, M. C., Rodriguez de Cordoba, S., Roumenina, L. T., Sethi, S., Smith, R. J. and Conference, P. (2017) Atypical hemolytic uremic syndrome and C3 glomerulopathy: conclusions from a "Kidney Disease: Improving Global Outcomes" (KDIGO) Controversies Conference. *Kidney Int.* **91**, 539-551
- Gore, S., Sanz Garcia, E., Hendrickx, P. M. S., Gutmanas, A., Westbrook, J. D., Yang, H., Feng, Z., Baskaran, K., Berrisford, J. M., Hudson, B. P., Ikegawa, Y., Kobayashi, N., Lawson, C. L., Mading, S., Mak, L., Mukhopadhyay, A., Oldfield, T. J., Patwardhan, A., Peisach, E., Sahni, G., Sekharan, M. R., Sen, S., Shao, C., Smart, O. S., Ulrich, E. L., Yamashita, R., Quesada, M., Young, J. Y., Nakamura, H., Markley, J. L., Berman, H. M., Burley, S. K., Velankar, S. and Kleywegt, G. J. (2017) Validation of Structures in the Protein Data Bank. *Structure.* **25**, 1916-1927
- Grant, A., Lee, D. and Orengo, C. (2004) Progress towards mapping the universe of protein folds. *Genome Biol.* **5**, 107
- Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon, J. A. and Caves, L. S. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics.* **22**, 2695-2696
- Griffiths, A. J. F. (2000) *An introduction to genetic analysis*, 7th ed. Ed., W.H. Freeman, New York
- Grumach, A. S. and Kirschfink, M. (2014) Are complement deficiencies really rare? Overview on prevalence, clinical importance and modern diagnostic approach. *Mol. Immunol.* **61**, 110-117
- Guinier, A., Fournet, G., Walker, C. B. and Yudowitch, K. L. (1955) *Small-angle scattering of X-rays*, Wiley ; London : Chapman and Hall, New York
- Guo, M. H., Dauber, A., Lippincott, M. F., Chan, Y. M., Salem, R. M. and Hirschhorn, J. N. (2016) Determinants of Power in Gene-Based Burden Testing for Monogenic Disorders. *Am. J. Hum. Genet.* **99**, 527-539
- Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C. and Shyr, Y. (2017) Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics.* **109**, 83-90
- Guttman, M., Weinkam, P., Sali, A. and Lee, K. K. (2013) All-atom ensemble modeling to analyze small-angle x-ray scattering of glycosylated proteins. *Structure.* **21**, 321-331
- Hageman, G. S., Anderson, D. H., Johnson, L. V., Hancox, L. S., Taiber, A. J., Hardisty, L. I., Hageman, J. L., Stockman, H. A., Borchardt, J. D., Gehrs, K. M., Smith, R. J., Silvestri, G., Russell, S. R., Klaver, C. C., Barbazetto, I., Chang, S., Yannuzzi, L. A., Barile, G. R., Merriam, J. C., Smith, R. T., Olsh, A. K., Bergeron, J., Zernant, J., Merriam, J. E., Gold, B., Dean, M. and Allikmets, R. (2005) A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7227-7232
- Hageman, G. S., Luthert, P. J., Chong, N. H. V., Johnson, L. V., Anderson, D. H. and Mullins, R. F. (2001) An integrated hypothesis that considers drusen as biomarkers of immune-mediated processes at the RPE-Bruch's membrane

- interface in aging and age-related macular degeneration. *Prog. Retin. Eye Res.* **20**, 705-732
- Hagler, A. T., Huler, E. and Lifson, S. (1974) Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Am. Chem. Soc.* **96**, 5319-5327
- Haimov, B. and Srebnik, S. (2016) A closer look into the alpha-helix basin. *Sci. Rep.* **6**, 38341
- Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., Spencer, K. L., Kwan, S. Y., Nouredine, M., Gilbert, J. R., Schnetz-Boutaud, N., Agarwal, A., Postel, E. A. and Pericak-Vance, M. A. (2005) Complement factor H variant increases the risk of age-related macular degeneration. *Science.* **308**, 419-421
- Hammel, M., Kriechbaum, M., Gries, A., Kostner, G. M., Laggner, P. and Prassl, R. (2002) Solution structure of human and bovine beta(2)-glycoprotein I revealed by small-angle X-ray scattering. *J. Mol. Biol.* **321**, 85-97
- Harrison, S. C. and Durbin, R. (1985) Is there a single pathway for the folding of a polypeptide chain? *Proc. Natl. Acad. Sci. U.S.A.* **82**, 4028
- Hecker, L. A., Edwards, A. O., Ryu, E., Tosakulwong, N., Baratz, K. H., Brown, W. L., Charbel Issa, P., Scholl, H. P., Pollok-Kopp, B., Schmid-Kubista, K. E., Bailey, K. R. and Oppermann, M. (2010) Genetic control of the alternative pathway of complement in humans and age-related macular degeneration. *Hum. Mol. Genet.* **19**, 209-215
- Hedrick, P. W. (2005) *Genetics of populations*, 3rd ed. Ed., Jones and Bartlett Publishers, Boston ; London
- Heeger, P. S., Lalli, P. N., Lin, F., Valujskikh, A., Liu, J., Muqim, N., Xu, Y. and Medof, M. E. (2005) Decay-accelerating factor modulates induction of T cell immunity. *J. Exp. Med.* **201**, 1523-1530
- Heinen, S., Hartmann, A., Lauer, N., Wiehl, U., Dahse, H. M., Schirmer, S., Gropp, K., Enghardt, T., Wallich, R., Halbach, S., Mihlan, M., Schlotzer-Schrehardt, U., Zipfel, P. F. and Skerka, C. (2009) Factor H-related protein 1 (CFHR-1) inhibits complement C5 convertase activity and terminal complex formation. *Blood.* **114**, 2439-2447
- Hellwage, J., Jokiranta, T. S., Koistinen, V., Vaarala, O., Meri, S. and Zipfel, P. F. (1999) Functional properties of complement factor H-related proteins FHR-3 and FHR-4: binding to the C3d region of C3b and differential regulation by heparin. *FEBS Lett.* **462**, 345-352
- Hellwage, J., Kuhn, S. and Zipfel, P. F. (1997) The human complement regulatory factor-H-like protein 1, which represents a truncated form of factor H, displays cell-attachment activity. *Biochem. J.* **326** (Pt 2), 321-327
- Helmy, K. Y., Katschke, K. J., Jr., Gorgani, N. N., Kljavin, N. M., Elliott, J. M., Diehl, L., Scales, S. J., Ghilardi, N. and van Lookeren Campagne, M. (2006) CRIg: a macrophage complement receptor required for phagocytosis of circulating pathogens. *Cell.* **124**, 915-927
- Henn, B. M., Botigue, L. R., Bustamante, C. D., Clark, A. G. and Gravel, S. (2015) Estimating the mutation load in human genomes. *Nat. Rev. Genet.* **16**, 333-343
- Henzler-Wildman, K. and Kern, D. (2007) Dynamic personalities of proteins. *Nature.* **450**, 964-972
- Herbert, A. P., Deakin, J. A., Schmidt, C. Q., Blaum, B. S., Egan, C., Ferreira, V. P., Pangburn, M. K., Lyon, M., Uhrin, D. and Barlow, P. N. (2007a) Structure shows that a glycosaminoglycan and protein recognition site in factor H is perturbed by age-related macular degeneration-linked single nucleotide polymorphism. *J. Biol. Chem.* **282**, 18960-18968

- Herbert, A. P., Deakin, J. A., Schmidt, C. Q., Blaum, B. S., Egan, C., Ferreira, V. P., Pangburn, M. K., Lyon, M., Uhrin, D. and Barlow, P. N. (2007b) Structure shows that a glycosaminoglycan and protein recognition site in factor H is perturbed by age-related macular degeneration-linked single nucleotide polymorphism. *J. Biol. Chem.* **282**, 18960-18968
- Hessing, M., Vlooswijk, R. A., Hackeng, T. M., Kanters, D. and Bouma, B. N. (1990) The localization of heparin-binding fragments on human C4b-binding protein. *J. Immunol.* **144**, 204-208
- Heurich, M., Martinez-Barricarte, R., Francis, N. J., Roberts, D. L., Rodriguez de Cordoba, S., Morgan, B. P. and Harris, C. L. (2011) Common polymorphisms in C3, factor B, and factor H collaborate to determine systemic complement activity and disease risk. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 8761-8766
- Heurich, M., Preston, R. J., O'Donnell, V. B., Morgan, B. P. and Collins, P. W. (2016) Thrombomodulin enhances complement regulation through strong affinity interactions with factor H and C3b-Factor H complex. *Thromb. Res.* **145**, 84-92
- Higgs, P. G. and Attwood, T. K. (2013) *Molecular Evolution and Population Genetics*, Wiley
- Hilbert, M., Bohm, G. and Jaenicke, R. (1993) Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins.* **17**, 138-151
- Hocking, H. G., Herbert, A. P., Kavanagh, D., Soares, D. C., Ferreira, V. P., Pangburn, M. K., Uhrin, D. and Barlow, P. N. (2008) Structure of the N-terminal region of complement factor H and conformational implications of disease-linked sequence variations. *J. Biol. Chem.* **283**, 9475-9487
- Holers, V. M. (2014) Complement and its receptors: new insights into human disease. *Annu. Rev. Immunol.* **32**, 433-459
- Holmberg, M. T., Blom, A. M. and Meri, S. (2001) Regulation of complement classical pathway by association of C4b-binding protein to the surfaces of SK-OV-3 and Caov-3 ovarian adenocarcinoma cells. *J. Immunol.* **167**, 935-939
- Hong-Wei, W. and Jia-Wei, W. (2017) How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Sci.* **26**, 32-39
- Hospital, A., Goñi, J. R., Orozco, M. and Gelpí, J. L. (2015) Molecular dynamics simulations: advances and applications. *Advances and Applications in Bioinformatics and Chemistry : AABC.* **8**, 37-47
- Hourcade, D., Holers, V. M. and Atkinson, J. P. (1989) The regulators of complement activation (RCA) gene cluster. *Adv. Immunol.* **45**, 381-416
- Hrdlickova, B., de Almeida, R. C., Borek, Z. and Withoff, S. (2014) Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochim. Biophys. Acta.* **1842**, 1910-1922
- Huang, I. K., Pei, J. and Grishin, N. V. (2013) Defining and predicting structurally conserved regions in protein superfamilies. *Bioinformatics.* **29**, 175-181
- Hubbard, R. E. and Kamran Haider, M. (2010) *Hydrogen Bonds in Proteins: Role and Strength*. in *eLS*, John Wiley and Sons Ltd, Chichester. pp
- Huerta, A., Arjona, E., Portoles, J., Lopez-Sanchez, P., Rabasco, C., Espinosa, M., Cavero, T., Blasco, M., Cao, M., Manrique, J., Cabello-Chavez, V., Suner, M., Heras, M., Fulladosa, X., Belmar, L., Sempere, A., Peralta, C., Castillo, L., Arnau, A., Praga, M. and Rodriguez de Cordoba, S. (2017) A retrospective study of pregnancy-associated atypical hemolytic uremic syndrome. *Kidney Int.*
- Hui, G. K., Wright, D. W., Vennard, O. L., Rayner, L. E., Pang, M., Yeo, S. C., Gor, J., Molyneux, K., Barratt, J. and Perkins, S. J. (2015) The solution structures of native and patient monomeric human IgA1 reveal asymmetric extended

- structures: implications for function and IgAN disease. *Biochem. J.* **471**, 167-185
- Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33-38, 27-38
- Iafate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949-951
- Iatropoulos, P., Noris, M., Mele, C., Piras, R., Valoti, E., Bresin, E., Curreri, M., Mondo, E., Zito, A., Gamba, S., Bettoni, S., Murer, L., Fremeaux-Bacchi, V., Vivarelli, M., Emma, F., Daina, E. and Remuzzi, G. (2016) Complement gene variants determine the risk of immunoglobulin-associated MPGN and C3 glomerulopathy and predict long-term renal outcome. *Mol. Immunol.* **71**, 131-142
- Inal, J. M., Hui, K. M., Miot, S., Lange, S., Ramirez, M. I., Schneider, B., Krueger, G. and Schifferli, J. A. (2005) Complement C2 receptor inhibitor trispanning: a novel human complement inhibitory receptor. *J. Immunol.* **174**, 356-366
- International HapMap, C., et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature.* **449**, 851-861
- Israelachvili, J. and Pashley, R. (1982) The hydrophobic interaction is long range, decaying exponentially with distance. *Nature.* **300**, 341-342
- Iwahara, J. and Clore, G. M. (2006) Detecting transient intermediates in macromolecular binding by paramagnetic NMR. *Nature.* **440**, 1227-1230
- Jager, R. D., Mieler, W. F. and Miller, J. W. (2008) Age-related macular degeneration. *N. Engl. J. Med.* **358**, 2606-2617
- Janeway, C. A. (2001) *Immunobiology 5 : the immune system in health and disease*, 5th ed. Ed., Garland ; Edinburgh : Churchill Livingstone, New York
- Janssen, B. J., Huizinga, E. G., Raaijmakers, H. C., Roos, A., Daha, M. R., Nilsson-Ekdahl, K., Nilsson, B. and Gros, P. (2005) Structures of complement component C3 provide insights into the function and evolution of immunity. *Nature.* **437**, 505-511
- Java, A., Liszewski, M. K., Hourcade, D. E., Zhang, F. and Atkinson, J. P. (2015) Role of complement receptor 1 (CR1; CD35) on epithelial cells: A model for understanding complement-mediated damage in the kidney. *Mol. Immunol.* **67**, 584-595
- Jo, S., Cheng, X., Islam, S. M., Huang, L., Rui, H., Zhu, A., Lee, H. S., Qi, Y., Han, W., Vanommeslaeghe, K., MacKerell, A. D., Jr., Roux, B. and Im, W. (2014) CHARMM-GUI PDB manipulator for advanced modeling and simulations of proteins containing nonstandard residues. *Adv. Protein Chem. Struct. Biol.* **96**, 235-265
- Jo, S., Kim, T., Iyer, V. G. and Im, W. (2008) CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859-1865
- Jo, S., Song, K. C., Desaire, H., MacKerell, A. D., Jr. and Im, W. (2011) Glycan Reader: automated sugar identification and simulation preparation for carbohydrates and glycoproteins. *J. Comput. Chem.* **32**, 3135-3141
- Jokiranta, T. S. (2017) HUS and atypical HUS. *Blood.* **129**, 2847-2856
- Jokiranta, T. S., Jaakola, V. P., Lehtinen, M. J., Parepalo, M., Meri, S. and Goldman, A. (2006) Structure of complement factor H carboxyl-terminus reveals molecular basis of atypical haemolytic uremic syndrome. *EMBO J.* **25**, 1784-1794
- Joosten, R. P., te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., Sander, C. and Vriend, G. (2011) A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **39**, D411-419

- Jorde, L. B. and Wooding, S. P. (2004) Genetic variation, classification and 'race'. *Nat. Genet.* **36**, S28-33
- Jozsi, M. and Zipfel, P. F. (2008) Factor H family proteins and human diseases. *Trends Immunol.* **29**, 380-387
- Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* **22**, 2577-2637
- Kajander, T., Lehtinen, M. J., Hyvarinen, S., Bhattacharjee, A., Leung, E., Isenman, D. E., Meri, S., Goldman, A. and Jokiranta, T. S. (2011) Dual interaction of factor H with C3d and glycosaminoglycans in host-nonhost discrimination by complement. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 2897-2902
- Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F. G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoeck, P., Tuli, M. A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R. (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **33**, D29-33
- Kapustin, Y., Chan, E., Sarkar, R., Wong, F., Vorechovsky, I., Winston, R. M., Tatusova, T. and Dibb, N. J. (2011) Cryptic splice sites and split genes. *Nucleic Acids Res.* **39**, 5837-5844
- Karki, R., Pandya, D., Elston, R. C. and Ferlini, C. (2015) Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC Med. Genomics.* **8**, 37
- Karplus, M. (1997) The Levinthal paradox: yesterday and today. *Fold Des.* **2**, S69-S75
- Karplus, M. (2011) Behind the folding funnel diagram. *Nat. Chem. Biol.* **7**, 401-404
- Karplus, M. and Kuriyan, J. (2005) Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6679-6685
- Karplus, M. and Weaver, D. L. (1994) Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci.* **3**, 650-668
- Katze, M. G., Korth, M. J., Law, G. L. and Nathanson, N. (2016) *Viral pathogenesis : from basics to systems biology*, 3rd Ed., Academic Press
- Kavanagh, D., Burgess, R., Spitzer, D., Richards, A., Diaz-Torres, M. L., Goodship, J. A., Hourcade, D. E., Atkinson, J. P. and Goodship, T. H. (2007) The decay accelerating factor mutation I197V found in hemolytic uraemic syndrome does not impair complement regulation. *Mol. Immunol.* **44**, 3162-3167
- Kavanagh, D., Richards, A. and Atkinson, J. (2008) Complement regulatory genes and hemolytic uremic syndromes. *Annu. Rev. Med.* **59**, 293-309
- Kearney, J. A. (2011) Genetic modifiers of neurological disease. *Curr. Opin. Genet. Dev.* **21**, 349-353
- Kelly, U., Yu, L., Kumar, P., Ding, J. D., Jiang, H., Hageman, G. S., Arshavsky, V. Y., Frank, M. M., Hauser, M. A. and Rickman, C. B. (2010) Heparan sulfate, including that in Bruch's membrane, inhibits the complement alternative pathway: implications for age-related macular degeneration. *J. Immunol.* **185**, 5486-5494
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res.* **12**, 996-1006
- Kerr, I. D., Cox, H. C., Moyes, K., Evans, B., Burdett, B. C., van Kan, A., McElroy, H., Vail, P. J., Brown, K. L., Sumampong, D. B., Monteferrante, N. J., Hardman, K. L., Theisen, A., Mundt, E., Wenstrup, R. J. and Eggington, J. M. (2017) Assessment of in silico protein sequence analysis in the clinical classification of variants in cancer risk genes. *J Community Genet.* **8**, 87-95

- Khan, S., Nan, R., Gor, J., Mulloy, B. and Perkins, S. J. (2012) Bivalent and co-operative binding of complement factor H to heparan sulfate and heparin. *Biochem. J.* **444**, 417-428
- Kim, H. Y. (2017) Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restor Dent Endod.* **42**, 152-155
- Kim, J. S., Afsari, B. and Chirikjian, G. S. (2017) Cross-Validation of Data Compatibility Between Small Angle X-ray Scattering and Cryo-Electron Microscopy. *J. Comput. Biol.* **24**, 13-30
- Kim, P. S. and Baldwin, R. L. (1982) Specific Intermediates in the Folding Reactions of Small Proteins and the Mechanism of Protein Folding. *Annu. Rev. Biochem.* **51**, 459-489
- Kimura, M. (1991) Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 5969-5973
- Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C. and Hoh, J. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science.* **308**, 385-389
- Kleijnung, J. and Fraternali, F. (2014) Design and application of implicit solvent models in biomolecular simulations. *Curr. Opin. Struct. Biol.* **25**, 126-134
- Klos, A., Tenner, A. J., Johswich, K. O., Ager, R. R., Reis, E. S. and Kohl, J. (2009) The role of the anaphylatoxins in health and disease. *Mol. Immunol.* **46**, 2753-2766
- Knopf, P. M., Rivera, D. S., Hai, S. H., McMurry, J., Martin, W. and De Groot, A. S. (2008) Novel function of complement C3d as an autologous helper T-cell target. *Immunol. Cell Biol.* **86**, 221-225
- Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S. E. and Topper, S. E. (2017) Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. *Genome Med.* **9**, 13
- Kodama, Y., Mashima, J., Kosuge, T., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y. and Takagi, T. (2018) DNA Data Bank of Japan: 30th anniversary. *Nucleic Acids Res.* **46**, D30-D35
- Kolodziejczyk, R., Mikula, K. M., Kotila, T., Postis, V. L. G., Jokiranta, T. S., Goldman, A. and Meri, T. (2017) Crystal structure of a tripartite complex between C3dg, C-terminal domains of factor H and OspE of *Borrelia burgdorferi*. *PLoS One.* **12**, e0188127
- Kouser, L., Abdul-Aziz, M., Nayak, A., Stover, C. M., Sim, R. B. and Kishore, U. (2013) Properdin and factor h: opposing players on the alternative complement pathway "see-saw". *Front. Immunol.* **4**, 93
- Krissinel, E. and Henrick, K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774-797
- Kruglyak, L. and Nickerson, D. A. (2001) Variation is the spice of life. *Nat. Genet.* **27**, 234-236
- Kryukov, G. V., Pennacchio, L. A. and Sunyaev, S. R. (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727-739
- Kuhn, S., Skerka, C. and Zipfel, P. F. (1995) Mapping of the complement regulatory domains in the human factor H-like protein 1 and in factor H1. *J. Immunol.* **155**, 5663-5670

- Kuhn, S. and Zipfel, P. F. (1996) Mapping of the domains required for decay acceleration activity of the human factor H-like protein 1 and factor H. *Eur. J. Immunol.* **26**, 2383-2387
- Kumar, N., Kapoor, A., Kalwar, A., Narayan, S., Singhal, M. K., Kumar, A., Mewara, A. and Bardia, M. R. (2014) Allele frequency of ABO blood group antigen and the risk of esophageal cancer. *Biomed Res Int.* **2014**, 286810
- Kumar, P., Henikoff, S. and Ng, P. C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073-1081
- Kumar, S. and Nussinov, R. (2002a) Close-range electrostatic interactions in proteins. *ChemBioChem.* **3**, 604-617
- Kumar, S. and Nussinov, R. (2002b) Relationship between ion pair geometries and electrostatic strengths in proteins. *Biophys. J.* **83**, 1595-1612
- Laine, M., Jarva, H., Seitsonen, S., Haapasalo, K., Lehtinen, M. J., Lindeman, N., Anderson, D. H., Johnson, P. T., Jarvela, I., Jokiranta, T. S., Hageman, G. S., Immonen, I. and Meri, S. (2007) Y402H polymorphism of complement factor H affects binding affinity to C-reactive protein. *J. Immunol.* **178**, 3831-3836
- Lambris, J. D., Lao, Z., Oglesby, T. J., Atkinson, J. P., Hack, C. E. and Becherer, J. D. (1996) Dissection of CR1, factor H, membrane cofactor protein, and factor B binding and functional sites in the third complement component. *J. Immunol.* **156**, 4821-4832
- Lambris, J. D. and Morikis, D. (2005) *Structural biology of the complement system*, 1st Ed., Taylor & Francis, Boca Raton, Florida
- Lander, E. S., et al. (2001) Initial sequencing and analysis of the human genome. *Nature.* **409**, 860-921
- Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., Kattman, B. L. and Maglott, D. R. (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062-D1067
- Langford-Smith, A., Keenan, T. D., Clark, S. J., Bishop, P. N. and Day, A. J. (2014) The role of complement in age-related macular degeneration: heparan sulphate, a ZIP code for complement factor H? *J. Innate Immun.* **6**, 407-416
- Lapeyraque, A. L., Fremeaux-Bacchi, V. and Robitaille, P. (2011) Efficacy of eculizumab in a patient with factor-H-associated atypical hemolytic uremic syndrome. *Pediatr. Nephrol.* **26**, 621-624
- Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., Paschall, J., Ananiev, V., Flicek, P. and Church, D. M. (2013) DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.* **41**, D936-941
- Law, S. K. and Dodds, A. W. (1997) The internal thioester and the covalent binding properties of the complement proteins C3 and C4. *Protein Sci.* **6**, 263-274
- Law, S. K., Lichtenberg, N. A., Holcombe, F. H. and Levine, R. P. (1980) Interaction between the labile binding sites of the fourth (C4) and fifth (C5) human complement proteins and erythrocyte cell membranes. *J. Immunol.* **125**, 634-639
- Law, S. K. A. and Reid, K. B. M. (1995) *Complement*, 2nd Ed., IRL Press at Oxford University Press, Oxford; New York
- Lee, B. and Richards, F. M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400
- Lee, J., Cheng, X., Swails, J. M., Yeom, M. S., Eastman, P. K., Lemkul, J. A., Wei, S., Buckner, J., Jeong, J. C., Qi, Y., Jo, S., Pande, V. S., Case, D. A., Brooks, C. L.,

- 3rd, MacKerell, A. D., Jr., Klauda, J. B. and Im, W. (2016) CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **12**, 405-413
- Lee, S., Abecasis, G. R., Boehnke, M. and Lin, X. (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5-23
- Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H. and Orengo, C. (2012) Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res.* **40**, D465-471
- Legendre, C. M., Licht, C., Muus, P., Greenbaum, L. A., Babu, S., Bedrosian, C., Bingham, C., Cohen, D. J., Delmas, Y., Douglas, K., Eitner, F., Feldkamp, T., Fouque, D., Furman, R. R., Gaber, O., Herthelius, M., Hourmant, M., Karpman, D., Lebranchu, Y., Mariat, C., Menne, J., Moulin, B., Nurnberger, J., Ogawa, M., Remuzzi, G., Richard, T., Sberro-Soussan, R., Severino, B., Sheerin, N. S., Trivelli, A., Zimmerhackl, L. B., Goodship, T. and Loirat, C. (2013) Terminal complement inhibitor eculizumab in atypical hemolytic-uremic syndrome. *N. Engl. J. Med.* **368**, 2169-2181
- Lek, M., et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* **536**, 285-291
- Lemaire, M., Fremeaux-Bacchi, V., Schaefer, F., Choi, M., Tang, W. H., Le Quintrec, M., Fakhouri, F., Taque, S., Nobili, F., Martinez, F., Ji, W., Overton, J. D., Mane, S. M., Nurnberg, G., Altmuller, J., Thiele, H., Morin, D., Deschenes, G., Baudouin, V., Llanas, B., Collard, L., Majid, M. A., Simkova, E., Nurnberg, P., Rioux-Leclerc, N., Moeckel, G. W., Gubler, M. C., Hwa, J., Loirat, C. and Lifton, R. P. (2013) Recessive mutations in DGKE cause atypical hemolytic-uremic syndrome. *Nat. Genet.* **45**, 531-536
- Lesser, G. J. and Rose, G. D. (1990) Hydrophobicity of amino acid subgroups in proteins. *Proteins.* **8**, 6-13
- Levinthal, C. (1968) Are there pathways of protein folding. *J. Chim. Phys.* **65**, 44-45
- Levy, M., Halbwachs-Mecarelli, L., Gubler, M. C., Kohout, G., Bensenouci, A., Niaudet, P., Hauptmann, G. and Lesavre, P. (1986) H deficiency in two brothers with atypical dense intramembranous deposit disease. *Kidney Int.* **30**, 949-956
- Levy, Y. and Onuchic, J. N. (2004) Water and proteins: a love-hate relationship. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 3325-3326
- Li, M., Atmaca-Sonmez, P., Othman, M., Branham, K. E., Khanna, R., Wade, M. S., Li, Y., Liang, L., Zarepari, S., Swaroop, A. and Abecasis, G. R. (2006) CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nat. Genet.* **38**, 1049-1054
- Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature.* **362**, 709-715
- Lindorfer, M. A., Pawluczkwycz, A. W., Peek, E. M., Hickman, K., Taylor, R. P. and Parker, C. J. (2010) A novel approach to preventing the hemolysis of paroxysmal nocturnal hemoglobinuria: both complement-mediated cytolysis and C3 deposition are blocked by a monoclonal antibody specific for the alternative pathway of complement. *Blood.* **115**, 2283-2291
- Lins, L., Thomas, A. and Brasseur, R. (2003) Analysis of accessible surface of residues in proteins. *Protein Sci.* **12**, 1406-1417
- Liszewski, M. K. and Atkinson, J. P. (2015) Complement regulator CD46: genetic variants and disease associations. *Hum Genomics.* **9**, 7
- Liszewski, M. K., Farries, T. C., Lublin, D. M., Rooney, I. A. and Atkinson, J. P. (1996) Control of the complement system. *Adv. Immunol.* **61**, 201-283

- Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. and Richardson, D. C. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins*. **50**, 437-450
- Loytynoja, A. and Goldman, N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 10557-10562
- Lubbers, R., van Essen, M. F., van Kooten, C. and Trouw, L. A. (2017) Production of complement components by cells of the immune system. *Clin. Exp. Immunol.* **188**, 183-194
- Ma, K. N., Cashman, S. M., Sweigard, J. H. and Kumar-Singh, R. (2010) Decay accelerating factor (CD55)-mediated attenuation of complement: therapeutic implications for age-related macular degeneration. *Invest. Ophthalmol. Vis. Sci.* **51**, 6776-6783
- MacArthur, D. G., et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. **335**, 823-828
- Maga, T. K., Meyer, N. C., Belsha, C., Nishimura, C. J., Zhang, Y. and Smith, R. J. (2011) A novel deletion in the RCA gene cluster causes atypical hemolytic uremic syndrome. *Nephrol. Dial. Transplant.* **26**, 739-741
- Maher, M. C., Uricchio, L. H., Torgerson, D. G. and Hernandez, R. D. (2012) Population genetics of rare variants and complex diseases. *Hum. Hered.* **74**, 118-128
- Mak, T. W., Saunders, M. E. and Jett, B. D. (2014) *Primer to the Immune Response*, 2nd Ed., Academic, Amsterdam ; London
- Makou, E., Mertens, H. D., Maciejewski, M., Soares, D. C., Matis, I., Schmidt, C. Q., Herbert, A. P., Svergun, D. I. and Barlow, P. N. (2012) Solution structure of CCP modules 10-12 illuminates functional architecture of the complement regulator, factor H. *J. Mol. Biol.* **424**, 295-312
- Maller, J., George, S., Purcell, S., Fagerness, J., Altshuler, D., Daly, M. J. and Seddon, J. M. (2006) Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat. Genet.* **38**, 1055-1059
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. and Visscher, P. M. (2009) Finding the missing heritability of complex diseases. *Nature*. **461**, 747-753
- Markwick, P. R., Malliavin, T. and Nilges, M. (2008) Structural biology by NMR: structure, dynamics, and interactions. *PLoS Comput. Biol.* **4**, e1000168
- Martinez-Barricarte, R., Heurich, M., Valdes-Canedo, F., Vazquez-Martul, E., Torreira, E., Montes, T., Tortajada, A., Pinto, S., Lopez-Trascasa, M., Morgan, B. P., Llorca, O., Harris, C. L. and Rodriguez de Cordoba, S. (2010) Human C3 mutation reveals a mechanism of dense deposit disease pathogenesis and provides insights into complement activation and regulation. *J. Clin. Invest.* **120**, 3702-3712
- Matsushita, M., Thiel, S., Jensenius, J. C., Terai, I. and Fujita, T. (2000) Proteolytic activities of two types of mannose-binding lectin-associated serine protease. *J. Immunol.* **165**, 2637-2642
- Matsuura, Y., Takehira, M., Joti, Y., Ogasahara, K., Tanaka, T., Ono, N., Kunishima, N. and Yutani, K. (2015) Thermodynamics of protein denaturation at temperatures over 100 °C: CutA1 mutant proteins substituted with hydrophobic and charged residues. *Sci. Rep.* **5**, 15545

- Matthews, K. W., Mueller-Ortiz, S. L. and Wetsel, R. A. (2004) Carboxypeptidase N: a pleiotropic regulator of inflammation. *Mol. Immunol.* **40**, 785-793
- McCammon, J. F., Gelin, B. R. and Karplus, M. (1977) Dynamics of folded proteins. *Nature.* **267**, 585-590
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P. and Cunningham, F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122
- Medicus, R. G., Gotze, O. and Muller-Eberhard, H. J. (1976) Alternative pathway of complement: recruitment of precursor properdin by the labile C3/C5 convertase and the potentiation of the pathway. *J. Exp. Med.* **144**, 1076-1093
- Medof, M. E., Iida, K., Mold, C. and Nussenzweig, V. (1982) Unique role of the complement receptor CR1 in the degradation of C3b associated with immune complexes. *J. Exp. Med.* **156**, 1739-1754
- Medof, M. E., Kinoshita, T. and Nussenzweig, V. (1984) Inhibition of complement activation on the surface of cells after incorporation of decay-accelerating factor (DAF) into their membranes. *J. Exp. Med.* **160**, 1558-1578
- Meri, S., Morgan, B. P., Wing, M., Jones, J., Davies, A., Podack, E. and Lachmann, P. J. (1990) Human protectin (CD59), an 18-20-kD homologous complement restriction factor, does not restrict perforin-mediated lysis. *J. Exp. Med.* **172**, 367-370
- Merinero, H. M., Garcia, S. P., Garcia-Fernandez, J., Arjona, E., Tortajada, A. and Rodriguez de Cordoba, S. (2017) Complete functional characterization of disease-associated genetic variants in the complement factor H gene. *Kidney Int.*
- Merle, N. S., Church, S. E., Fremeaux-Bacchi, V. and Roumenina, L. T. (2015a) Complement System Part I - Molecular Mechanisms of Activation and Regulation. *Front. Immunol.* **6**, 262
- Merle, N. S., Noe, R., Halbwachs-Mecarelli, L., Fremeaux-Bacchi, V. and Roumenina, L. T. (2015b) Complement System Part II: Role in Immunity. *Front. Immunol.* **6**, 257
- Mertens, H. D. T. and Svergun, D. I. (2017) Combining NMR and small angle X-ray scattering for the study of biomolecular structure and dynamics. *Arch. Biochem. Biophys.* **628**, 33-41
- Milder, F. J., Gomes, L., Schouten, A., Janssen, B. J., Huizinga, E. G., Romijn, R. A., Hemrika, W., Roos, A., Daha, M. R. and Gros, P. (2007) Factor B structure provides insights into activation of the central protease of the complement system. *Nat. Struct. Mol. Biol.* **14**, 224-228
- Min, L., Nie, M., Zhang, A., Wen, J., Noel, S. D., Lee, V., Carroll, R. S. and Kaiser, U. B. (2016) Computational Analysis of Missense Variants of G Protein-Coupled Receptors Involved in the Neuroendocrine Regulation of Reproduction. *Neuroendocrinology.* **103**, 230-239
- Mitchell, K. J. (2012) What is complex about complex disorders? *Genome Biol.* **13**, 237
- Montes, T., Tortajada, A., Morgan, B. P., Rodriguez de Cordoba, S. and Harris, C. L. (2009) Functional basis of protection against age-related macular degeneration conferred by a common polymorphism in complement factor B. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 4366-4371
- Morgan, B. P. (1990) *Complement : clinical aspects and relevance to disease*, Academic Press
- Morgan, B. P. and Gasque, P. (1997) Extrahepatic complement biosynthesis: where, when and why? *Clin. Exp. Immunol.* **107**, 1-7
- Morgan, H. P., Mertens, H. D., Guariento, M., Schmidt, C. Q., Soares, D. C., Svergun, D. I., Herbert, A. P., Barlow, P. N. and Hannan, J. P. (2012) Structural analysis

- of the C-terminal region (modules 18-20) of complement regulator factor H (FH). *PLoS One*. **7**, e32187
- Morgan, H. P., Schmidt, C. Q., Guariento, M., Blaum, B. S., Gillespie, D., Herbert, A. P., Kavanagh, D., Mertens, H. D., Svergun, D. I., Johansson, C. M., Uhrin, D., Barlow, P. N. and Hannan, J. P. (2011) Structural basis for engagement by complement factor H of C3b on a self surface. *Nat. Struct. Mol. Biol.* **18**, 463-470
- Motulsky, A. G. (2006) Genetics of complex diseases. *J. Zhejiang Univ. Sci. B*. **7**, 167-168
- Munoz, J. and Heck, A. J. (2014) From the human genome to the human proteome. *Angew. Chem. Int. Ed. Engl.* **53**, 10864-10866
- Murata, K. and Wolf, M. (2018) Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochim. Biophys. Acta*. **1862**, 324-334
- Myers, J. K., Pace, C. N. and Scholtz, J. M. (1997) Helix propensities are identical in proteins and peptides. *Biochemistry*. **36**, 10923-10929
- Myles, S., Davison, D., Barrett, J., Stoneking, M. and Timpson, N. (2008) Worldwide population differentiation at disease-associated SNPs. *BMC Med. Genomics*. **1**, 22
- Nan, R., Gor, J., Lengyel, I. and Perkins, S. J. (2008b) Uncontrolled zinc- and copper-induced oligomerisation of the human complement regulator factor H and its possible implications for function and disease. *J. Mol. Biol.* **384**, 1341-1352
- Nan, R., Gor, J. and Perkins, S. J. (2008a) Implications of the progressive self-association of wild-type human factor H for complement regulation and disease. *J. Mol. Biol.* **375**, 891-900
- Nan, R., Ward, G., Gavigan, L., Miller, A., Gor, J., Lengyel, I. and Perkins, S. J. (2010) The His402 allotype of complement factor H show similar self-association to the Tyr402 allotype but exhibits greater self-association in the presence of zinc. *Mol. Immunol.* **47**, 2263
- Narayan, M. (2012) Disulfide bonds: protein folding and subcellular protein trafficking. *FEBS J.* **279**, 2272-2282
- Nauta, A. J., Trouw, L. A., Daha, M. R., Tijsma, O., Nieuwland, R., Schwaeble, W. J., Gingras, A. R., Mantovani, A., Hack, E. C. and Roos, A. (2002) Direct binding of C1q to apoptotic cells and cell blebs induces complement activation. *Eur. J. Immunol.* **32**, 1726-1736
- Nei, M. and Nozawa, M. (2011) Roles of mutation and selection in speciation: from Hugo de Vries to the modern genomic era. *Genome Biol. Evol.* **3**, 812-829
- Nelson, R. A., Jr., Jensen, J., Gigli, I. and Tamura, N. (1966) Methods for the separation, purification and measurement of nine components of hemolytic complement in guinea-pig serum. *Immunochemistry*. **3**, 111-135
- Nesargikar, P. N., Spiller, B. and Chavez, R. (2012) The complement system: history, pathways, cascade and inhibitors. *Eur. J. Microbiol. Immunol. (Bp)*. **2**, 103-111
- Nester, C. M., Barbour, T., de Cordoba, S. R., Dragon-Durey, M. A., Fremeaux-Bacchi, V., Goodship, T. H., Kavanagh, D., Noris, M., Pickering, M., Sanchez-Corral, P., Skerka, C., Zipfel, P. and Smith, R. J. (2015) Atypical aHUS: State of the art. *Mol. Immunol.* **67**, 31-42
- Neyman, J. and Pearson, E. S. (1933) IX. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. London, Ser. A*. **231**, 289
- Nicol, P. A. and Lachmann, P. J. (1973) The alternate pathway of complement activation. The role of C3 and its inactivator (KAF). *Immunology*. **24**, 259-275
- Nilsson, S. C., Sim, R. B., Lea, S. M., Fremeaux-Bacchi, V. and Blom, A. M. (2011) Complement factor I in health and disease. *Mol. Immunol.* **48**, 1611-1620

- Nogales, E. and Scheres, S. H. (2015) Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. *Mol. Cell.* **58**, 677-689
- Nonaka, M. and Kimura, A. (2006) Genomic view of the evolution of the complement system. *Immunogenetics.* **58**, 701-713
- Noris, M. and Remuzzi, G. (2009) Atypical hemolytic-uremic syndrome. *N. Engl. J. Med.* **361**, 1676-1687
- Noris, M. and Remuzzi, G. (2015) Glomerular diseases dependent on complement activation, including atypical hemolytic uremic syndrome, membranoproliferative glomerulonephritis, and C3 glomerulopathy: core curriculum 2015. *Am. J. Kidney Dis.* **66**, 359-375
- Noris, M. and Remuzzi, G. (2017) Genetics of Immune-Mediated Glomerular Diseases: Focus on Complement. *Semin. Nephrol.* **37**, 447-463
- O'Leary, N. A., et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745
- Okemefuna, A. I., Gilbert, H. E., Griggs, K. M., Ormsby, R. J., Gordon, D. L. and Perkins, S. J. (2008) The regulatory SCR-1/5 and cell surface-binding SCR-16/20 fragments of factor H reveal partially folded-back solution structures and different self-associative properties. *J. Mol. Biol.* **375**, 80-101
- Okemefuna, A. I., Nan, R., Gor, J. and Perkins, S. J. (2009) Electrostatic interactions contribute to the folded-back conformation of wild type human factor H. *J. Mol. Biol.* **391**, 98-118
- Okemefuna, A. I., Nan, R., Miller, A., Gor, J. and Perkins, S. J. (2010) Complement factor H binds at two independent sites to C-reactive protein in acute phase concentrations. *J. Biol. Chem.* **285**, 1053-1065
- Onuchic, J. N. and Wolynes, P. G. (2004) Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**, 70-75
- Orengo, C. A., Jones, D. and Thornton, J. M. (2003) *Bioinformatics : genes, proteins and computers*, Bios, Oxford
- Ormsby, R. J., Jokiranta, T. S., Duthy, T. G., Griggs, K. M., Sadlon, T. A., Giannakis, E. and Gordon, D. L. (2006) Localization of the third heparin-binding site in the human complement regulator factor H1. *Mol. Immunol.* **43**, 1624-1632
- Ormsby, R. J., Ranganathan, S., Tong, J. C., Griggs, K. M., Dimasi, D. P., Hewitt, A. W., Burdon, K. P., Craig, J. E., Hoh, J. and Gordon, D. L. (2008) Functional and structural implications of the complement factor H Y402H polymorphism associated with age-related macular degeneration. *Invest. Ophthalmol. Vis. Sci.* **49**, 1763-1770
- Ortega, A., Amoros, D. and Garcia de la Torre, J. (2011) Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models. *Biophys. J.* **101**, 892-898
- Osborne, A. J., Breno, M., Borsa, N. G., Bu, F., Fremeaux-Bacchi, V., Gale, D. P., van den Heuvel, L. P., Kavanagh, D., Noris, M., Pinto, S., Rallapalli, P. M., Remuzzi, G., Rodriguez de Cordoba, S., Ruiz, A., Smith, R. J. H., Vieira-Martins, P., Volokhina, E., Wilson, V., Goodship, T. H. J. and Perkins, S. J. (2018a) Statistical Validation of Rare Complement Variants Provides Insights into the Molecular Basis of Atypical Hemolytic Uremic Syndrome and C3 Glomerulopathy. *J. Immunol.* **200**, 2464-2478
- Osborne, A. J., Nan, R., Miller, A., Bhatt, J., Gor, J. and Perkins, S. J. (2018b) Two distinct conformations of factor H regulate discrete complement-binding functions in the fluid phase and at cell surfaces. *J. Biol. Chem.* **293**, 17166-17187

- Pace, C. N. and Scholtz, J. M. (1998) A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* **75**, 422-427
- Pan, J., Zhang, L., Organtini, L. J., Hafenstein, S. and Bergelson, J. M. (2015) Specificity of coxsackievirus B3 interaction with human, but not murine, decay-accelerating factor: replacement of a single residue within short consensus repeat 2 prevents virus attachment. *J. Virol.* **89**, 1324-1328
- Pangburn, M. K. (2000) Host recognition and target differentiation by factor H, a regulator of the alternative pathway of complement. *Immunopharmacology.* **49**, 149-157
- Pangburn, M. K. and Muller-Eberhard, H. J. (1983) Initiation of the alternative complement pathway due to spontaneous hydrolysis of the thioester of C3. *Ann. N. Y. Acad. Sci.* **421**, 291-298
- Pangburn, M. K. and Muller-Eberhard, H. J. (1986) The C3 convertase of the alternative pathway of human complement. Enzymic properties of the bimolecular proteinase. *Biochem. J.* **235**, 723-730
- Pangburn, M. K., Rawal, N., Cortes, C., Alam, M. N., Ferreira, V. P. and Atkinson, M. A. (2009) Polyanion-induced self-association of complement factor H. *J. Immunol.* **182**, 1061-1068
- Pangburn, M. K., Schreiber, R. D. and Muller-Eberhard, H. J. (1977) Human complement C3b inactivator: isolation, characterization, and demonstration of an absolute requirement for the serum protein beta1H for cleavage of C3b and C4b in solution. *J. Exp. Med.* **146**, 257-270
- Park, C. T. and Wright, S. D. (1996) Plasma lipopolysaccharide-binding protein is found associated with a particle containing apolipoprotein A-I, phospholipid, and factor H-related proteins. *J. Biol. Chem.* **271**, 18054-18060
- Park, H. J., Guariento, M., Maciejewski, M., Hauhart, R., Tham, W. H., Cowman, A. F., Schmidt, C. Q., Mertens, H. D., Liszewski, M. K., Hourcade, D. E., Barlow, P. N. and Atkinson, J. P. (2014) Using mutagenesis and structural biology to map the binding site for the Plasmodium falciparum merozoite protein PfRh4 on the human immune adherence receptor. *J. Biol. Chem.* **289**, 450-463
- Pelikan, M., Hura, G. L. and Hammel, M. (2009) Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen. Physiol. Biophys.* **28**, 174-189
- Pennisi, E. (2012) Genomics. ENCODE project writes eulogy for junk DNA. *Science.* **337**, 1159, 1161
- Perkins, S. J. (1986) Protein volumes and hydration effects. The calculations of partial specific volumes, neutron scattering matchpoints and 280-nm absorption coefficients for proteins and glycoproteins from amino acid sequences. *Eur. J. Biochem.* **157**, 169-180
- Perkins, S. J., Fung, K. W. and Khan, S. (2014) Molecular Interactions between Complement Factor H and Its Heparin and Heparan Sulfate Ligands. *Front. Immunol.* **5**, 126
- Perkins, S. J., Nan, R., Li, K., Khan, S. and Abe, Y. (2011) Analytical ultracentrifugation combined with X-ray and neutron scattering: Experiment and modelling. *Methods.* **54**, 181-199
- Perkins, S. J., Nan, R., Li, K., Khan, S. and Miller, A. (2012) Complement factor H-ligand interactions: self-association, multivalency and dissociation constants. *Immunobiology.* **217**, 281-297
- Perkins, S. J., Nan, R., Okemefuna, A. I., Li, K., Khan, S. and Miller, A. (2010a) Multiple interactions of complement Factor H with its ligands in solution: a progress report. *Adv. Exp. Med. Biol.* **703**, 25-47

- Perkins, S. J., Nealis, A. S. and Sim, R. B. (1991) Oligomeric domain structure of human complement factor H by X-ray and neutron solution scattering. *Biochemistry*. **30**, 2847-2857
- Perkins, S. J., Okemefuna, A. I., Fernando, A. N., Bonner, A., Gilbert, H. E. and Furtado, P. B. (2008) X-ray and neutron scattering data and their constrained molecular modelling. *Methods Cell Biol.* **84**, 375-423
- Perkins, S. J., Okemefuna, A. I. and Nan, R. (2010b) Unravelling protein-protein interactions between complement factor H and C-reactive protein using a multidisciplinary strategy. *Biochem. Soc. Trans.* **38**, 894-900
- Perkins, S. J., Okemefuna, A. I., Nan, R., Li, K. and Bonner, A. (2009) Constrained solution scattering modelling of human antibodies and complement proteins reveals novel biological insights. *J. R. Soc. Interface.* **6 Suppl 5**, S679-696
- Perkins, S. J., Wright, D. W., Zhang, H. L., Brookes, E. H., Chen, J. H., Irving, T. C., Krueger, S., Barlow, D. J., Edler, K. J., Scott, D. J., Terrill, N. J., King, S. M., Butler, P. D. and Curtis, J. E. (2016) Atomistic modelling of scattering data in the Collaborative Computational Project for Small Angle Scattering (CCP-SAS). *J. Appl. Crystallogr.* **49**, 1861-1875
- Persson, B. D., Schmitz, N. B., Santiago, C., Zocher, G., Larvie, M., Scheu, U., Casasnovas, J. M. and Stehle, T. (2010) Structure of the extracellular portion of CD46 provides insights into its interactions with complement proteins and pathogens. *PLoS Pathog.* **6**, e1001122
- Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D., Konarev, P. V. and Svergun, D. I. (2012) New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45**, 342-350
- Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. and Goldstein, D. B. (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709
- Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L. and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781-1802
- Pickering, M. C., Cook, H. T., Warren, J., Bygrave, A. E., Moss, J., Walport, M. J. and Botto, M. (2002) Uncontrolled C3 activation causes membranoproliferative glomerulonephritis in mice deficient in complement factor H. *Nat. Genet.* **31**, 424-428
- Pickering, M. C., de Jorge, E. G., Martinez-Barricarte, R., Recalde, S., Garcia-Layana, A., Rose, K. L., Moss, J., Walport, M. J., Cook, H. T., de Cordoba, S. R. and Botto, M. (2007) Spontaneous hemolytic uremic syndrome triggered by complement factor H lacking surface recognition domains. *J. Exp. Med.* **204**, 1249-1256
- Ponnuraj, K., Xu, Y., Macon, K., Moore, D., Volanakis, J. E. and Narayana, S. V. (2004) Structural analysis of engineered Bb fragment of complement factor B: insights into the activation mechanism of the alternative pathway C3-convertase. *Mol. Cell.* **14**, 17-28
- Pray, L. A. (2008) DNA Replication and Causes of Mutation. *Nature Education.* **1**, 214
- Pross, S. (2007) *T-Cell Activation*. in *xPharm: The Comprehensive Pharmacology Reference*, Elsevier, New York. pp 1-7
- Prosser, B. E., Johnson, S., Roversi, P., Clark, S. J., Tarelli, E., Sim, R. B., Day, A. J. and Lea, S. M. (2007a) Expression, purification, cocrystallization and preliminary crystallographic analysis of sucrose octasulfate/human complement regulator factor H SCRs 6-8. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **63**, 480-483

- Prosser, B. E., Johnson, S., Roversi, P., Herbert, A. P., Blaum, B. S., Tyrrell, J., Jowitt, T. A., Clark, S. J., Tarelli, E., Uhrin, D., Barlow, P. N., Sim, R. B., Day, A. J. and Lea, S. M. (2007b) Structural basis for complement factor H linked age-related macular degeneration. *J. Exp. Med.* **204**, 2277-2283
- Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61-65
- R Core Team. (2013) R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*
- Rallapalli, P. M. (2014) Interactive locus-specific databases and evolutionary aspects of the mutations in coagulation proteins. *PhD Thesis, University College London.*
- Rallapalli, P. M., Kembball-Cook, G., Tuddenham, E. G., Gomez, K. and Perkins, S. J. (2013) An interactive mutation database for human coagulation factor IX provides novel insights into the phenotypes and genetics of hemophilia B. *J. Thromb. Haemost.* **11**, 1329-1340
- Rallapalli, P. M., Orengo, C. A., Studer, R. A. and Perkins, S. J. (2014) Positive selection during the evolution of the blood coagulation factors in the context of their disease-causing mutations. *Mol. Biol. Evol.* **31**, 3040-3056
- Ramachandran, G. N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95-99
- Ramachandran, G. N. and Sasisekharan, V. (1968) Conformation of polypeptides and proteins. *Adv. Protein Chem.* **23**, 283-438
- Raskatov, J. A. and Teplow, D. B. (2017) Using chirality to probe the conformational dynamics and assembly of intrinsically disordered amyloid proteins. *Sci. Rep.* **7**, 12433
- Raychaudhuri, S., Iartchouk, O., Chin, K., Tan, P. L., Tai, A. K., Ripke, S., Gowrisankar, S., Vemuri, S., Montgomery, K., Yu, Y., Reynolds, R., Zack, D. J., Campochiaro, B., Campochiaro, P., Katsanis, N., Daly, M. J. and Seddon, J. M. (2011) A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. *Nat. Genet.* **43**, 1232-1236
- Receveur-Brechot, V. and Durand, D. (2012) How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr Protein Pept Sci.* **13**, 55-75
- Reeves, G. A., Talavera, D. and Thornton, J. M. (2009) Genome and proteome annotation: organization, interpretation and integration. *J. R. Soc. Interface.* **6**, 129-147
- Reif, M. M., Winger, M. and Oostenbrink, C. (2013) Testing of the GROMOS Force-Field Parameter Set 54A8: Structural Properties of Electrolyte Solutions, Lipid Bilayers, and Proteins. *J. Chem. Theory Comput.* **9**, 1247-1264
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L. and Committee, A. L. Q. A. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405-424
- Ricklin, D., Reis, E. S. and Lambris, J. D. (2016) Complement in disease: a defence system turning offensive. *Nat. Rev. Nephrol.* **12**, 383-401
- Rigden, D. J. (2009) *From protein structure to function with bioinformatics*, Springer, Dordrecht
- Ripoche, J., Day, A. J., Harris, T. J. and Sim, R. B. (1988) The complete amino acid sequence of human complement factor H. *Biochem. J.* **249**, 593-602
- Robinson, M. A. (1998) *Linkage Disequilibrium A2 - Delves, Peter J. in Encyclopedia of Immunology (Second Edition)*, Elsevier, Oxford. pp 1586-1588

- Rodriguez de Cordoba, S., Esparza-Gordillo, J., Goicoechea de Jorge, E., Lopez-Trascasa, M. and Sanchez-Corral, P. (2004) The human complement factor H: functional roles, genetic variations and disease associations. *Mol. Immunol.* **41**, 355-367
- Rodriguez de Cordoba, S. and Goicoechea de Jorge, E. (2008) Translational mini-review series on complement factor H: genetics and disease associations of human complement factor H. *Clin. Exp. Immunol.* **151**, 1-13
- Rodriguez, E., Rallapalli, P. M., Osborne, A. J. and Perkins, S. J. (2014) New functional and structural insights from updated mutational databases for complement factor H, Factor I, membrane cofactor protein and C3. *Biosci. Rep.* **34**, 635-649
- Rose, P. W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., Costanzo, L. D., Duarte, J. M., Dutta, S., Feng, Z., Green, R. K., Goodsell, D. S., Hudson, B., Kalro, T., Lowe, R., Peisach, E., Randle, C., Rose, A. S., Shao, C., Tao, Y. P., Valasatava, Y., Voigt, M., Westbrook, J. D., Woo, J., Yang, H., Young, J. Y., Zardecki, C., Berman, H. M. and Burley, S. K. (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45**, D271-D281
- Rosenberg, N. A. and Kang, J. T. (2015) Genetic Diversity and Societally Important Disparities. *Genetics.* **201**, 1-12
- Ross, G. D., Newman, S. L., Lambris, J. D., Devery-Pocius, J. E., Cain, J. A. and Lachmann, P. J. (1983) Generation of three different fragments of bound C3 with purified factor I or serum. II. Location of binding sites in the C3 fragments for factors B and H, complement receptors, and bovine conglutinin. *J. Exp. Med.* **158**, 334-352
- Roumenina, L. T., Frimat, M., Miller, E. C., Provot, F., Dragon-Durey, M. A., Bordereau, P., Bigot, S., Hue, C., Satchell, S. C., Mathieson, P. W., Mousson, C., Noel, C., Sautes-Fridman, C., Halbwachs-Mecarelli, L., Atkinson, J. P., Lionet, A. and Fremeaux-Bacchi, V. (2012) A prevalent C3 mutation in aHUS patients causes a direct C3 convertase gain of function. *Blood.* **119**, 4182-4191
- Roumenina, L. T., Jablonski, M., Hue, C., Blouin, J., Dimitrov, J. D., Dragon-Durey, M. A., Cayla, M., Fridman, W. H., Macher, M. A., Ribes, D., Moulonguet, L., Rostaing, L., Satchell, S. C., Mathieson, P. W., Sautes-Fridman, C., Loirat, C., Regnier, C. H., Halbwachs-Mecarelli, L. and Fremeaux-Bacchi, V. (2009) Hyperfunctional C3 convertase leads to complement deposition on endothelial cells and contributes to atypical hemolytic uremic syndrome. *Blood.* **114**, 2837-2845
- Rousset, F. and Raymond, M. (1995) Testing heterozygote excess and deficiency. *Genetics.* **140**, 1413-1419
- Roversi, P., Johnson, S., Caesar, J. J., McLean, F., Leath, K. J., Tsiftoglou, S. A., Morgan, B. P., Harris, C. L., Sim, R. B. and Lea, S. M. (2011) Structural basis for complement factor I control and its disease-associated sequence polymorphisms. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 12839-12844
- Roy, A., Kucukural, A. and Zhang, Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725-738
- Ruggenenti, P., Noris, M. and Remuzzi, G. (2001) Thrombotic microangiopathy, hemolytic uremic syndrome, and thrombotic thrombocytopenic purpura. *Kidney Int.* **60**, 831-846
- Ruvinsky, A. M., Kirys, T., Tuzikov, A. V. and Vakser, I. A. (2012) Structure fluctuations and conformational changes in protein binding. *J. Bioinform. Comput. Biol.* **10**, 1241002

- Sadallah, S., Gudat, F., Laissue, J. A., Spath, P. J. and Schifferli, J. A. (1999) Glomerulonephritis in a patient with complement factor I deficiency. *Am. J. Kidney Dis.* **33**, 1153-1157
- Saibil, H. R. (1996) What can electron microscopy tell us about chaperoned protein folding? *Fold Des.* **1**, R45-49
- Sali, A. and Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815
- Sali, A., Shakhnovich, E. and Karplus, M. (1994) How does a protein fold? *Nature.* **369**, 248-251
- Sanchez Chinchilla, D., Pinto, S., Hoppe, B., Adragna, M., Lopez, L., Justa Roldan, M. L., Pena, A., Lopez Trascasa, M., Sanchez-Corral, P. and Rodriguez de Cordoba, S. (2014) Complement mutations in diacylglycerol kinase-epsilon-associated atypical hemolytic uremic syndrome. *Clin. J. Am. Soc. Nephrol.* **9**, 1611-1619
- Sanders, M. F. and Bowman, J. L. (2012) *Genetic analysis : an integrated approach*, 1st Ed., Benjamin Cummings, Boston, Mass. ; London
- Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463-5467
- Sansbury, F. H., Cordell, H. J., Bingham, C., Bromilow, G., Nicholls, A., Powell, R., Shields, B., Smyth, L., Warwicker, P., Strain, L., Wilson, V., Goodship, J. A., Goodship, T. H. and Turnpenny, P. D. (2014) Factors determining penetrance in familial atypical haemolytic uraemic syndrome. *J. Med. Genet.* **51**, 756-764
- Sassa, A., Kanemaru, Y., Kamoshita, N., Honma, M. and Yasui, M. (2016) Mutagenic consequences of cytosine alterations site-specifically embedded in the human genome. *Genes Environ.* **38**, 17
- Saunders, R. E., Abarrategui-Garrido, C., Fremeaux-Bacchi, V., Goicoechea de Jorge, E., Goodship, T. H., Lopez Trascasa, M., Noris, M., Ponce Castro, I. M., Remuzzi, G., Rodriguez de Cordoba, S., Sanchez-Corral, P., Skerka, C., Zipfel, P. F. and Perkins, S. J. (2007) The interactive Factor H-atypical hemolytic uremic syndrome mutation database and website: update and integration of membrane cofactor protein and Factor I mutations with structural models. *Hum. Mutat.* **28**, 222-234
- Saunders, R. E., Goodship, T. H., Zipfel, P. F. and Perkins, S. J. (2006) An interactive web database of factor H-associated hemolytic uremic syndrome mutations: insights into the structural consequences of disease-associated mutations. *Hum. Mutat.* **27**, 21-30
- Saunders, R. E. and Perkins, S. J. (2006) A user's guide to the interactive Web database of factor H-associated hemolytic uremic syndrome. *Semin. Thromb. Hemost.* **32**, 160-168
- Scheuner, M. T., Yoon, P. W. and Khoury, M. J. (2004) Contribution of Mendelian disorders to common chronic disease: opportunities for recognition, intervention, and prevention. *Am. J. Med. Genet. C Semin. Med. Genet.* **125C**, 50-65
- Schmidt, B., Ho, L. and Hogg, P. J. (2006) Allosteric disulfide bonds. *Biochemistry.* **45**, 7429-7433
- Schmidt, B. Z., Fowler, N. L., Hidvegi, T., Perlmutter, D. H. and Colten, H. R. (1999) Disruption of disulfide bonds is responsible for impaired secretion in human complement factor H deficiency. *J. Biol. Chem.* **274**, 11782-11788
- Schmidt, C. Q., Bai, H., Lin, Z., Risitano, A. M., Barlow, P. N., Ricklin, D. and Lambris, J. D. (2013) Rational engineering of a minimized immune inhibitor with unique triple-targeting properties. *J. Immunol.* **190**, 5712-5721

- Schmidt, C. Q., Herbert, A. P., Kavanagh, D., Gandy, C., Fenton, C. J., Blaum, B. S., Lyon, M., Uhrin, D. and Barlow, P. N. (2008) A new map of glycosaminoglycan and C3b binding sites on factor H. *J. Immunol.* **181**, 2610-2619
- Schmidt, C. Q., Herbert, A. P., Mertens, H. D., Guariento, M., Soares, D. C., Uhrin, D., Rowe, A. J., Svergun, D. I. and Barlow, P. N. (2010) The central portion of factor H (modules 10-15) is compact and contains a structurally deviant CCP module. *J. Mol. Biol.* **395**, 105-122
- Schnell, A. H. and Witte, J. S. (2008) *Molecular epidemiology : applications in cancer and other human diseases*, 1st Ed., Informa Healthcare, New York
- Schramm, E. C., Roumenina, L. T., Rybkine, T., Chauvet, S., Vieira-Martins, P., Hue, C., Maga, T., Valoti, E., Wilson, V., Jokiranta, S., Smith, R. J., Noris, M., Goodship, T., Atkinson, J. P. and Fremeaux-Bacchi, V. (2015) Mapping interactions between complement C3 and regulators using mutations in atypical hemolytic uremic syndrome. *Blood.* **125**, 2359-2369
- Schrodinger, LLC. (2015) *The PyMOL Molecular Graphics System, Version 1.8*,
- Schwaeble, W. J. and Reid, K. B. (1999) Does properdin crosslink the cellular and the humoral immune response? *Immunol. Today.* **20**, 17-21
- Schwendinger, M. G., Spruth, M., Schoch, J., Dierich, M. P. and Proding, W. M. (1997) A novel mechanism of alternative pathway complement activation accounts for the deposition of C3 fragments on CR2-expressing homologous cells. *J. Immunol.* **158**, 5455-5463
- Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A. and Wigler, M. (2004) Large-scale copy number polymorphism in the human genome. *Science.* **305**, 525-528
- Sellier-Leclerc, A. L., Fremeaux-Bacchi, V., Dragon-Durey, M. A., Macher, M. A., Niaudet, P., Guest, G., Boudailliez, B., Bouissou, F., Deschenes, G., Gie, S., Tsimaratos, M., Fischbach, M., Morin, D., Nivet, H., Alberti, C., Loirat, C. and French Society of Pediatric, N. (2007) Differential impact of complement mutations on clinical characteristics in atypical hemolytic uremic syndrome. *J. Am. Soc. Nephrol.* **18**, 2392-2400
- Semenyuk, A. V. and Svergun, D. I. (1991) Gnom - a Program Package for Small-Angle Scattering Data-Processing. *J. Appl. Crystallogr.* **24**, 537-540
- Servais, A., Noel, L. H., Roumenina, L. T., Le Quintrec, M., Ngo, S., Dragon-Durey, M. A., Macher, M. A., Zuber, J., Karras, A., Provot, F., Moulin, B., Grunfeld, J. P., Niaudet, P., Lesavre, P. and Fremeaux-Bacchi, V. (2012) Acquired and genetic complement abnormalities play a critical role in dense deposit disease and other C3 glomerulopathies. *Kidney Int.* **82**, 454-464
- Seya, T., Holers, V. M. and Atkinson, J. P. (1985) Purification and functional analysis of the polymorphic variants of the C3b/C4b receptor (CR1) and comparison with H, C4b-binding protein (C4bp), and decay accelerating factor (DAF). *J. Immunol.* **135**, 2661-2667
- Shafee, T. and Lowe, R. (2017) Eukaryotic and prokaryotic gene structure. *WikiJournal of Medicine.* **4**, 2
- Sham, P. C. and Purcell, S. M. (2014) Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335-346
- Sharma, A. K. and Pangburn, M. K. (1996) Identification of three physically and functionally distinct binding sites for C3b in human complement factor H by deletion mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10996-11001
- Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson,

- D. G., Pinkel, D. and Eichler, E. E. (2005) Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78-88
- Shen, M. Y. and Sali, A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507-2524
- Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A. and Waterston, R. H. (2017) DNA sequencing at 40: past, present and future. *Nature.* **550**, 345-353
- Shiang, R., Murray, J. C., Morton, C. C., Buetow, K. H., Wasmuth, J. J., Olney, A. H., Sanger, W. G. and Goldberger, G. (1989) Mapping of the human complement factor I gene to 4q25. *Genomics.* **4**, 82-86
- Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W. Z., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D. and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**
- Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., Lehtinen, S., Studer, R. A., Thornton, J. and Orengo, C. A. (2015) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* **43**, D376-381
- Sim, R. B. and Reboul, A. (1981) Preparation and properties of human C1 inhibitor. *Methods Enzymol.* **80 Pt C**, 43-54
- Sim, R. B., Twose, T. M., Paterson, D. S. and Sim, E. (1981) The covalent-binding reaction of complement component C3. *Biochem. J.* **193**, 115-127
- Simons, K. T., Kooperberg, C., Huang, E. and Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209-225
- Skerka, C., Chen, Q., Fremeaux-Bacchi, V. and Roumenina, L. T. (2013) Complement factor H related proteins (CFHRs). *Mol. Immunol.* **56**, 170-180
- Skerka, C., Lauer, N., Weinberger, A. A., Keilhauer, C. N., Suhnel, J., Smith, R., Schlotzer-Schrehardt, U., Fritsche, L., Heinen, S., Hartmann, A., Weber, B. H. and Zipfel, P. F. (2007) Defective complement control of factor H (Y402H) and FHL-1 in age-related macular degeneration. *Mol. Immunol.* **44**, 3398-3406
- Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C. and Froguel, P. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature.* **445**, 881-885
- Smith, B. O., Mallin, R. L., Krych-Goldberg, M., Wang, X., Hauhart, R. E., Bromek, K., Uhrin, D., Atkinson, J. P. and Barlow, P. N. (2002) Structure of the C3b binding site of CR1 (CD35), the immune adherence receptor. *Cell.* **108**, 769-780
- Smith, K. A. (2012) Louis Pasteur, the father of immunology? *Front. Immunol.* **3**, 68
- Smith, R. J., Harris, C. L. and Pickering, M. C. (2011) Dense deposit disease. *Mol. Immunol.* **48**, 1604-1610
- Song, J. J., Hwang, I., Cho, K. H., Garcia, M. A., Kim, A. J., Wang, T. H., Lindstrom, T. M., Lee, A. T., Nishimura, T., Zhao, L., Morser, J., Nesheim, M., Goodman, S. B., Lee, D. M., Bridges, S. L., Jr., Consortium for the Longitudinal Evaluation of African Americans with Early Rheumatoid Arthritis, R., Gregersen, P. K., Leung, L. L. and Robinson, W. H. (2011) Plasma carboxypeptidase B downregulates inflammatory responses in autoimmune arthritis. *J. Clin. Invest.* **121**, 3517-3527
- Song, W., Gardner, S. A., Hovhannisyan, H., Natalizio, A., Weymouth, K. S., Chen, W., Thibodeau, I., Bogdanova, E., Letovsky, S., Willis, A. and Nagan, N. (2016)

- Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. *Genet. Med.* **18**, 850-854
- Stahl, A. L., Kristoffersson, A., Olin, A. I., Olsson, M. L., Roodhooft, A. M., Proesmans, W. and Karpman, D. (2009) A novel mutation in the complement regulator clusterin in recurrent hemolytic uremic syndrome. *Mol. Immunol.* **46**, 2236-2243
- Strohmeyer, R., Ramirez, M., Cole, G. J., Mueller, K. and Rogers, J. (2002) Association of factor H of the alternative pathway of complement with agrin and complement receptor 3 in the Alzheimer's disease brain. *J. Neuroimmunol.* **131**, 135-146
- Sullivan, M., Erlic, Z., Hoffmann, M. M., Arbeiter, K., Patzer, L., Budde, K., Hoppe, B., Zeier, M., Lhotta, K., Rybicki, L. A., Bock, A., Berisha, G. and Neumann, H. P. (2010) Epidemiological approach to identifying genetic predispositions for atypical hemolytic uremic syndrome. *Ann. Hum. Genet.* **74**, 17-26
- Sun, Y., Ruivenkamp, C. A., Hoffer, M. J., Vrijenhoek, T., Kriek, M., van Asperen, C. J., den Dunnen, J. T. and Santen, G. W. (2015) Next-generation diagnostics: gene panel, exome, or whole genome? *Hum. Mutat.* **36**, 648-655
- Sun, Z., Reid, K. B. and Perkins, S. J. (2004) The dimeric and trimeric solution structures of the multidomain complement protein properdin by X-ray scattering, analytical ultracentrifugation and constrained modelling. *J. Mol. Biol.* **343**, 1327-1343
- Tabor, H. K., Risch, N. J. and Myers, R. M. (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat. Rev. Genet.* **3**, 391-397
- Tan, L. A., Yu, B., Sim, F. C., Kishore, U. and Sim, R. B. (2010) Complement activation by phospholipids: the interplay of factor H and C1q. *Protein Cell.* **1**, 1033-1049
- Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., Akey, J. M., Broad, G. O., Seattle, G. O. and Project, N. E. S. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science.* **337**, 64-69
- Thakkinstian, A., Han, P., McEvoy, M., Smith, W., Hoh, J., Magnusson, K., Zhang, K. and Attia, J. (2006) Systematic review and meta-analysis of the association between complement factor H Y402H polymorphisms and age-related macular degeneration. *Hum. Mol. Genet.* **15**, 2784-2790
- The UniProt, C. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158-D169
- Thomas, P. J., Qu, B. H. and Pedersen, P. L. (1995) Defective protein folding as a basis of human disease. *Trends Biochem. Sci.* **20**, 456-459
- Thomas, S., Ranganathan, D., Francis, L., Madhan, K. and John, G. T. (2014) Current concepts in C3 glomerulopathy. *Indian J. Nephrol.* **24**, 339-348
- Thompson, R. A. and Winterborn, M. H. (1981) Hypocomplementaemia due to a genetic deficiency of beta 1H globulin. *Clin. Exp. Immunol.* **46**, 110-119
- Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. and Wilke, C. O. (2013) Maximum allowed solvent accessibilities of residues in proteins. *PLoS One.* **8**, e80635
- Timmann, C., Leippe, M. and Horstmann, R. D. (1991) Two major serum components antigenically related to complement factor H are different glycosylation forms of

- a single protein with no factor H-like complement regulatory functions. *J. Immunol.* **146**, 1265-1270
- Todd, A. E., Orengo, C. A. and Thornton, J. M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113-1143
- Toomey, C. B., Johnson, L. V. and Bowes Rickman, C. (2018) Complement factor H in AMD: Bridging genetic associations and pathobiology. *Prog. Retin. Eye Res.* **62**, 38-57
- Tortajada, A., Montes, T., Martinez-Barricarte, R., Morgan, B. P., Harris, C. L. and de Cordoba, S. R. (2009) The disease-protective complement factor H allotypic variant Ile62 shows increased binding affinity for C3b and enhanced cofactor activity. *Hum. Mol. Genet.* **18**, 3452-3461
- Tortajada, A., Yebenes, H., Abarrategui-Garrido, C., Anter, J., Garcia-Fernandez, J. M., Martinez-Barricarte, R., Alba-Dominguez, M., Malik, T. H., Bedoya, R., Cabrera Perez, R., Lopez Trascasa, M., Pickering, M. C., Harris, C. L., Sanchez-Corral, P., Llorca, O. and Rodriguez de Cordoba, S. (2013) C3 glomerulopathy-associated CFHR1 mutation alters FHR oligomerization and complement regulation. *J. Clin. Invest.* **123**, 2434-2446
- Trivedi, M. V., Laurence, J. S. and Siahhaan, T. J. (2009) The role of thiols and disulfides on protein stability. *Curr. Protein Peptide Sci.* **10**, 614-625
- Uversky, V. N. and Dunker, A. K. (2010) Understanding protein non-folding. *Biochim. Biophys. Acta.* **1804**, 1231-1264
- Valoti, E., Alberti, M., Tortajada, A., Garcia-Fernandez, J., Gastoldi, S., Besso, L., Bresin, E., Remuzzi, G., Rodriguez de Cordoba, S. and Noris, M. (2015) A novel atypical hemolytic uremic syndrome-associated hybrid CFHR1/CFH gene encoding a fusion protein that antagonizes factor H-dependent complement regulation. *J. Am. Soc. Nephrol.* **26**, 209-219
- van den Heuvel, L., Riesbeck, K., El Tahir, O., Gracchi, V., Kremlitzka, M., Morre, S. A., van Furth, A. M., Singh, B., Okroj, M., van de Kar, N., Blom, A. M. and Volokhina, E. (2018) Genetic predisposition to infection in a case of atypical hemolytic uremic syndrome. *J. Hum. Genet.* **63**, 93-96
- van der Sijde, M. R., Ng, A. and Fu, J. (2014) Systems genetics: From GWAS to disease pathways. *Biochim. Biophys. Acta.* **1842**, 1903-1909
- Venables, J. P., Strain, L., Routledge, D., Bourn, D., Powell, H. M., Warwicker, P., Diaz-Torres, M. L., Sampson, A., Mead, P., Webb, M., Pirson, Y., Jackson, M. S., Hughes, A., Wood, K. M., Goodship, J. A. and Goodship, T. H. (2006) Atypical haemolytic uraemic syndrome associated with a hybrid complement gene. *PLoS Med.* **3**, e431
- Vieira-Martins, P., El Sissy, C., Bordereau, P., Gruber, A., Rosain, J. and Fremeaux-Bacchi, V. (2016) Defining the genetics of thrombotic microangiopathies. *Transfus. Apher. Sci.* **54**, 212-219
- Villarrreal, S. A. and Stewart, P. L. (2014) CryoEM and image sorting for flexible protein/DNA complexes. *J. Struct. Biol.* **187**, 76-83
- Vyse, T. J., Bates, G. P., Walport, M. J. and Morley, B. J. (1994) The organization of the human complement factor I gene (IF): a member of the serine protease gene family. *Genomics.* **24**, 90-98
- Wagner, E. K., Raychaudhuri, S., Villalonga, M. B., Java, A., Triebwasser, M. P., Daly, M. J., Atkinson, J. P. and Seddon, J. M. (2016) Mapping rare, deleterious mutations in Factor H: Association with early onset, drusen burden, and lower antigenic levels in familial AMD. *Sci. Rep.* **6**, 31531
- Wagner, M. J. (2013) Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. *Pharmacogenomics.* **14**, 413-424

- Wakeley, J. (1996) The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends Ecol. Evol.* **11**, 158-162
- Walport, M. J. (2002) Complement and systemic lupus erythematosus. *Arthritis Res.* **4 Suppl 3**, S279-293
- Walsh, R., Thomson, K. L., Ware, J. S., Funke, B. H., Woodley, J., McGuire, K. J., Mazzarotto, F., Blair, E., Seller, A., Taylor, J. C., Minikel, E. V., Exome Aggregation, C., MacArthur, D. G., Farrall, M., Cook, S. A. and Watkins, H. (2016) Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet. Med.*
- Wang, H., Vinnikov, I., Shahzad, K., Bock, F., Ranjan, S., Wolter, J., Kashif, M., Oh, J., Bierhaus, A., Nawroth, P., Kirschfink, M., Conway, E. M., Madhusudhan, T. and Isermann, B. (2012) The lectin-like domain of thrombomodulin ameliorates diabetic glomerulopathy via complement inhibition. *Thromb. Haemost.* **108**, 1141-1153
- Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. and Jones, D. T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635-645
- Warwicker, P., Goodship, T. H., Donne, R. L., Pirson, Y., Nicholls, A., Ward, R. M., Turnpenny, P. and Goodship, J. A. (1998) Genetic studies into inherited and sporadic hemolytic uremic syndrome. *Kidney Int.* **53**, 836-844
- Watson, J. D. and Crick, F. H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature.* **171**, 737-738
- Watson, M. C. and Curtis, J. E. (2013) Rapid and accurate calculation of small-angle scattering profiles using the golden ratio. *J. Appl. Crystallogr.* **46**, 1171-1177
- Webb, B. and Sali, A. (2016) Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Protein Sci.* **86**, 29 1-29 37
- Webb, J. H., Blom, A. M. and Dahlback, B. (2002) Vitamin K-dependent protein S localizing complement regulator C4b-binding protein to the surface of apoptotic cells. *J. Immunol.* **169**, 2580-2586
- Weiler, J. M., Daha, M. R., Austen, K. F. and Fearon, D. T. (1976) Control of the amplification convertase of complement by the plasma protein beta1H. *Proc. Natl. Acad. Sci. U.S.A.* **73**, 3268-3272
- Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P. (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765-784
- Weiss, A. J. R. and Scott. (2009) Chapter 20 - Epidemiologic and Population Genetic Studies. in *Clin. Transl. Sci.*, Academic Press, San Diego. pp 289-299
- Welling, L. and Thomson, L. (2009) *PHP and MySQL Web development*, 4th ed. Ed., Addison-Wesley, 2008, Upper Saddle River, NJ ; London
- Welte, T., Arnold, F., Kappes, J., Seidl, M., Haffner, K., Bergmann, C., Walz, G. and Neumann-Haefelin, E. (2018) Treating C3 glomerulopathy with eculizumab. *BMC Nephrol.* **19**, 7
- Wetlaufer, D. B. (1973) Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **70**, 697
- Whaley, K. and Ruddy, S. (1976) Modulation of the alternative complement pathways by beta 1 H globulin. *J. Exp. Med.* **144**, 1147-1163
- Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A.,

- Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L. and Yaschenko, E. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**, D13-21
- Wiita, A. P., Ainavarapu, S. R., Huang, H. H. and Fernandez, J. M. (2006) Force-dependent chemical kinetics of disulfide bond reduction observed with single-molecule techniques. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 7222-7227
- Williams, H. E. and Lane, D. (2002) *Web database applications with PHP and MySQL*, O'Reilly, Beijing ; Cambridge
- Wittke-Thompson, J. K., Pluzhnikov, A. and Cox, N. J. (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 967-986
- Wright, D. W. and Perkins, S. J. (2015) SCT: a suite of programs for comparing atomistic models with small-angle scattering data. *J. Appl. Crystallogr.* **48**, 953-961
- Wu, J., Wu, Y. Q., Ricklin, D., Janssen, B. J., Lambris, J. D. and Gros, P. (2009) Structure of complement fragment C3b-factor H and implications for host protection by complement regulators. *Nat. Immunol.* **10**, 728-733
- Wüthrich, K. (1986) *NMR of proteins and nucleic acids*, Wiley, New York ; Chichester
- Xue, X., Wu, J., Ricklin, D., Forneris, F., Di Crescenzo, P., Schmidt, C. Q., Granneman, J., Sharp, T. H., Lambris, J. D. and Gros, P. (2017) Regulator-dependent mechanisms of C3b processing by factor I allow differentiation of immune responses. *Nat. Struct. Mol. Biol.* **24**, 643-651
- Yang, L. W., Eyal, E., Bahar, I. and Kitao, A. (2009) Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics. *Bioinformatics.* **25**, 606-614
- Yang, S. (2014) Methods for SAXS-based Topological Structure Determination of Biomolecular Complexes. *Adv. Mater.* **26**, 7902-7910
- Zetterberg, M., Landgren, S., Andersson, M. E., Palmer, M. S., Gustafson, D. R., Skoog, I., Minthon, L., Thelle, D. S., Wallin, A., Bogdanovic, N., Andreasen, N., Blennow, K. and Zetterberg, H. (2008) Association of complement factor H Y402H gene polymorphism with Alzheimer's disease. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147B**, 720-726
- Zhang, W., Howell, S. C., Wright, D. W., Heindel, A., Qiu, X., Chen, J. and Curtis, J. E. (2017) Combined Monte Carlo/torsion-angle molecular dynamics for ensemble modeling of proteins, nucleic acids and carbohydrates. *J. Mol. Graph. Model.* **73**, 179-190
- Zhu, Q., Ge, D., Maia, J. M., Zhu, M., Petrovski, S., Dickson, S. P., Heinzen, E. L., Shianna, K. V. and Goldstein, D. B. (2011) A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *Am. J. Hum. Genet.* **88**, 458-468
- Zondervan, K. T. and Cardon, L. R. (2004) The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5**, 89-100
- Zondervan, K. T. and Cardon, L. R. (2007) Designing candidate gene and genome-wide case-control association studies. *Nat. Protoc.* **2**, 2492-2501
- Zuber, J., Fakhouri, F., Roumenina, L. T., Loirat, C., Fremeaux-Bacchi, V. and French Study Group for a, H. C. G. (2012) Use of eculizumab for atypical haemolytic uraemic syndrome and C3 glomerulopathies. *Nat. Rev. Nephrol.* **8**, 643-657
- Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R. and Lander, E. S. (2014) Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E455-464

Appendix I. Rare variants classified as ‘benign’ or ‘likely benign’ with their allele frequency and functional details (continues overleaf)

Gene	Variant	Protein level	Allele frequency (%)					Function reference	<i>In silico</i> : PolyPhen-2	<i>In silico</i> : SIFT	Class
			aHUS	C3G	ExAC	1000GP	EVS				
C3	c.-3_-2dupAC	-	0.02	0.13	0.88	0.22	0	-	-	-	Likely benign
C3	c.463A>C	p.Lys155Gln	0.37	0.13	0.34	0.04	0.29	-	Benign, neutral	Tolerated, neutral	Likely benign
C3	c.696C>T	p.Phe232Phe	0.02	0	0	0	0	-	-	Tolerated, neutral	Likely benign
C3	c.975C>T	p.Tyr325Tyr	0.02	0	0	0	0	-	-	Tolerated, neutral	Likely benign
C3	c.1146T>C	p.Ser382Ser	0	0.26	0.001	0	0	-	-	Tolerated, neutral	Likely benign
C3	c.1650C>T	p.Ser550Ser	0	0.13	0.01	0	0.02	-	-	Tolerated, neutral	Likely benign
C3	c.1758G>A	p.Glu586Glu	0.02	0	0	0	0	-	-	Tolerated, neutral	Likely benign
C3	c.2190C>T	p.Tyr730Tyr	0	0.26	0.04	0.04	0.05	-	-	Tolerated, neutral	Likely benign
C3	c.2203C>T	p.Arg735Trp	0.22	0.26	0.21	0.08	0.26	Fremaux-Bacchi et al, 2008	Probably damaging, deleterious	Damaging, deleterious	Benign
C3	c.2901C>T	p.Leu967Leu	0.02	0	0.20	0.08	0.19	Clinvar	-	Tolerated, neutral	Likely benign
C3	c.3671G>A	p.Gly1224Asp	0.04	0	0.18	0.76	0.70	Clinvar	Benign, neutral	Tolerated, deleterious	Likely benign
C3	c.4645C>A	p.Leu1549Met	0.02	0	0.12	0.40	0.34	Clinvar	Possibly damaging, deleterious	Damaging, neutral	Likely benign
C3	c.4827C>T	p.Ser1609Ser	0.02	0	0.01	0	0.01	-	-	Tolerated, neutral	Likely benign
C3	c.4855A>C	p.Ser1619Arg	0.20	0.13	0.11	0.04	0.22	Clinvar	Possibly damaging, deleterious	Tolerated, neutral	Likely benign
CD46	c.38C>T	p.Ser13Phe	0.09	0	0.50	0.86	0.15	-	Probably damaging, deleterious	Tolerated, neutral	Likely benign
CD46	c.66G>A	p.Ala22Ala	0	0.12	0	0	0	-	-	Tolerated, neutral	Likely benign
CD46	c.796G>A	p.Asp266Asn	0.03	0	0.12	0.44	0.42	-	Benign, neutral	Tolerated, neutral	Likely benign
CFB	c.1524C>T	p.His508His	0.04	0	0.31	0.18	0.20	-	-	Tolerated, neutral	Likely benign
CFH	c.172T>G	p.Ser58Ala	0.06	0	0.01	0.02	0.02	Merinero et al, 2017	Benign, neutral	Tolerated, neutral	Benign
CFH	c.849A>G	p.Lys283Lys	0.02	0	0	0	0	-	-	Tolerated, neutral	Likely benign
CFH	c.1707C>T	p.Cys569Cys	0.02	0	0.04	0.06	0.19	-	-	Tolerated, neutral	Likely benign
CFH	c.1824C>T	p.Ser608Ser	0.02	0	0.01	0.02	0.01	-	-	Tolerated, neutral	Likely benign
CFH	c.1949G>T	p.Gly650Val	0.02	0.11	0.02	0.04	0.02	-	Benign, neutral	Tolerated, neutral	Likely benign
CFH	c.2468T>C	p.Met823Thr	0.02	0	0	0	0	Merinero et al, 2017	Possibly damaging, deleterious	Damaging, deleterious	Benign
CFH	c.2509G>A	p.Val837Ile	0.02	0	0.13	0.34	0	Clinvar	Benign, neutral	Tolerated, neutral	Likely benign
CFH	c.2850G>T	p.Gln950His	0.37	0.45	0.36	0.16	0.44	Clinvar	Possibly damaging, deleterious	Damaging, deleterious	Likely benign
CFH	c.2867C>T	p.Thr956Met	0.14	0.56	0.12	0.12	0.12	Merinero et al, 2017	Probably damaging, deleterious	Tolerated, neutral	Benign
CFH	c.3050C>T	p.Thr1017Ile	0.03	0.34	0.12	0.30	0.52	Clinvar	Possibly damaging, deleterious	Damaging, neutral	Likely benign
CFH	c.3172T>C	p.Tyr1058His	0.29	0.11	0	0.14	0	-	Benign, neutral	Tolerated, neutral	Likely benign
CFH	c.3178G>C	p.Val1060Leu	0.26	0	0	0	0	-	Benign, neutral	Tolerated, neutral	Likely benign
CFH	c.3207T>C	p.Ser1069Ser	0	0.11	0.10	0.18	0.17	-	-	Tolerated, neutral	Likely benign
CFHR5	c.329T>C	p.Val110Ala	0.16	0	0.21	0.38	0.01	-	-	Damaging, deleterious	Likely benign

Appendix I. (continued) Rare variants classified as ‘benign’ or ‘likely benign’ with their allele frequency and functional details

Gene	Variant	Protein level	Allele frequency (%)					Function reference	<i>In silico</i> : PolyPhen-2	<i>In silico</i> : SIFT	Class
			aHUS	C3G	ExAC	1000GP	EVS				
<i>CFHR5</i>	c.432A>T	p.Lys144Asn	0.16	0	0.12	0.12	0.07	-	-	Tolerated, neutral	Likely benign
<i>CFHR5</i>	c.832G>A	p.Gly278Ser	0.32	0.48	0.73	0.36	0.69	-	-	Damaging, deleterious	Likely benign
<i>CFHR5</i>	c.1586T>G	p.Leu529Arg	0	0.96	0.11	0.26	0.51	-	-	Tolerated, deleterious	Likely benign
<i>CFI</i>	c.608C>T	p.Thr203Ile	0.05	0.12	0.05	0.04	0.05	-	Benign, neutral	Tolerated, neutral	Likely benign
<i>CFI</i>	c.782G>A	p.Gly261Asp	0.22	0	0.13	0.04	0.17	Fremaux-Bacchi et al, 2013	Benign, neutral	Tolerated, neutral	Likely benign
<i>CFI</i>	c.916A>G	p.Ile306Val	0.03	0	0.05	0.12	0.15	-	Benign, neutral	Tolerated, neutral	Likely benign
<i>CFI</i>	c.1322A>G	p.Lys441Arg	0.21	0.37	0.34	0.12	0.36	Clinvar	Benign, neutral	Tolerated, neutral	Likely benign
<i>CFI</i>	c.1657C>T	p.Pro553Ser	0.15	0.12	0.13	0.04	0.14	Fremaux-Bacchi et al, 2013	Benign, neutral	Tolerated, deleterious	Likely benign
<i>CFI</i>	c.*144T>C	-	0.02	0	0.27	0	0	Clinvar	-	-	Likely benign
<i>CFI</i>	c.*7G>T	-	0	0.12	0.27	0	0.78	-	-	-	Likely benign
<i>DGKE</i>	c.35C>T	p.Pro12Leu	0.21	0	0.49	0.20	0.56	-	Benign, neutral	Damaging, neutral	Likely benign
<i>DGKE</i>	c.1679A>G	p.Gln560Arg	0	0.39	0.20	0.08	0.18	-	Benign, neutral	Tolerated, neutral	Likely benign
<i>PLG</i>	c.112A>G	p.Lys38Glu	0	0.48	0.33	0.22	0.43	-	-	Damaging, neutral	Likely benign
<i>PLG</i>	c.266G>A	p.Arg89Lys	0.35	0.48	0.66	0.18	0.80	-	Benign, neutral	Tolerated, neutral	Likely benign
<i>PLG</i>	c.1259G>A	p.Gly420Asp	0	0.96	0.19	0.08	0.15	-	-	Tolerated, neutral	Likely benign
<i>PLG</i>	c.1469G>A	p.Arg490Gln	0.52	0.48	0.12	0.08	0.19	-	Probably damaging, deleterious	Damaging, deleterious	Likely benign
<i>PLG</i>	c.1481C>T	p.Ala494Val	0.70	0	0.82	0.90	0.85	-	Probably damaging, deleterious	Damaging, neutral	Likely benign
<i>THBD</i>	c.40G>A	p.Gly14Ser	0.04	0.19	0.22	0.68	0.37	Clinvar	Benign, neutral	Tolerated, neutral	Likely benign
<i>THBD</i>	c.939C>A	p.Gly313Gly	0	0.19	0	0.02	0	-	-	Tolerated, neutral	Likely benign
<i>THBD</i>	c.1437C>T	p.Asp479Asp	0.04	0	0	0	0	-	-	Tolerated, neutral	Likely benign
<i>THBD</i>	c.1504G>C	p.Gly502Arg	0.04	0	0.23	0.76	0.62	-	Possibly damaging, deleterious	Tolerated, neutral	Likely benign
<i>THBD</i>	c.1536C>G	p.Leu512Leu	0.04	0	0	0	0	-	-	Tolerated, neutral	Likely benign

Appendix II. The frequency of individuals with a rare variant (RV) out of all individuals screened per gene in the aHUS, C3G and ExAC datasets using an AF cut off of 0.01%

Gene	Variant effect	aHUS					ExAC					C3G			
		RVs	RV cases	All cases	Frequency (%)	<i>P</i> ^a	RVs	RV cases	All cases	Frequency (%)	<i>P</i> ^b	RVs	RV cases	All cases	Frequency (%)
<i>CFH</i>	Protein-altering	193	411	3128	13.14	<0.0001 ^c	342	734	59163	1.24	<0.0001 ^c	22	23	443	5.19
<i>C3</i>	Protein-altering	64	177	2455	7.21	<0.0001 ^c	448	923	59545	1.55	<0.0001 ^c	24	26	379	6.86
<i>CD46</i>	Protein-altering	74	219	2942	7.44	<0.0001 ^c	136	240	58338	0.41	0.18	0	0	406	0.00
<i>CFI</i>	Protein-altering	68	98	2923	3.35	<0.0001 ^c	197	382	60289	0.63	0.48	2	2	408	0.49
<i>DGKE</i>	Protein-altering	19	21	703	2.99	<0.0001 ^c	174	273	58439	0.47	0.45	0	0	127	0.00
<i>THBD</i>	Protein-altering	6	7	1328	0.53	0.37	144	247	38602	0.64	0.09	3	4	264	1.52
<i>CFB</i>	Protein-altering	15	22	2457	0.90	0.09	215	382	58732	0.65	0.10	6	5	379	1.32
<i>PLG</i>	Protein-altering	2	2	286	0.70	0.49	251	536	60619	0.88	0.28	2	2	104	1.92
<i>CFHR5</i>	Protein-altering	2	2	317	0.63	0.42	254	535	59751	0.90	0.50	1	1	104	0.96
<i>CFH</i>	Non-truncating ^e	129	318	3128	10.17	<0.0001 ^c	334	725	59286	1.22	<0.0001 ^c	17	19	443	4.29
<i>C3</i>	Non-truncating	59	171	2455	6.97	<0.0001 ^c	436	901	59566	1.51	<0.0001 ^c	20	21	379	5.54
<i>CFI</i>	Non-truncating	55	83	2923	2.84	<0.0001 ^c	179	353	60259	0.59	0.50	2	2	408	0.49
<i>CD46</i>	Non-truncating	43	99	2942	3.37	<0.0001 ^c	124	214	58634	0.37	0.21	0	0	406	0.00
<i>DGKE</i>	Non-truncating	7	9	703	1.28	<0.006 ^c	159	247	58385	0.42	0.48	0	0	127	0.00
<i>CFHR5</i>	Non-truncating	1	1	317	0.32	0.30	211	438	59650	0.73	0.50	1	1	104	0.96
<i>THBD</i>	Non-truncating	6	7	1328	0.53	0.39	138	239	37880	0.63	0.08	3	4	264	1.52
<i>CFB</i>	Non-truncating	15	22	2457	0.90	0.06	197	364	58710	0.62	0.23	5	4	379	1.06
<i>PLG</i>	Non-truncating	2	2	286	0.70	0.50	236	519	60632	0.86	0.26	2	2	104	1.92
<i>CD46</i>	Truncating ^f	31	125	2942	4.25	<0.0001 ^c	12	26	55275	0.05	0.50	0	0	406	0.00
<i>CFH</i>	Truncating	64	98	3128	3.13	<0.0001 ^c	8	9	54027	0.02	<0.0001 ^c	5	5	443	1.13
<i>DGKE</i>	Truncating	12	18	703	2.56	<0.0001 ^c	15	26	59004	0.04	0.50	0	0	127	0.00
<i>CFI</i>	Truncating	13	16	2923	0.55	<0.0001 ^c	18	29	60583	0.05	0.50	0	0	408	0.00
<i>C3</i>	Truncating	5	6	2455	0.24	<0.0001 ^c	12	22	58798	0.04	<0.0001 ^c	4	5	379	1.32
<i>THBD</i>	Truncating	0	0	1328	0.00	0.50	6	8	55211	0.02	0.50	0	0	264	0.00
<i>CFB</i>	Truncating	0	0	2457	0.00	0.40	18	18	58973	0.03	<0.0001 ^c	1	1	379	0.26
<i>PLG</i>	Truncating	0	0	286	0.00	0.50	15	17	60408	0.03	0.50	0	0	104	0.00
<i>CFHR5</i>	Truncating	1	1	317	0.32	0.50	43	97	60244	0.16	0.50	0	0	104	0.00

^a p value after the Chi-Square test (χ^2) with Yates' correction (p<0.05 after Bonferroni correction) for aHUS and ExAC

^b p value after the Chi-Square test (χ^2) with Yates' correction (p<0.05 after Bonferroni correction) for C3G and ExAC

^c Denotes aHUS or C3G dataset frequency significantly greater than ExAC frequency

^d Denotes ExAC frequency significantly greater than aHUS or C3G dataset frequency

^e Defined as missense or in-frame variant

^f Defined as nonsense, frameshift, or splice acceptor/donor variant

Appendix III. The frequency of individuals with a rare variant (RVs) out of all individuals screened per gene in the aHUS, C3G and EVS datasets using an AF cut off of 0.01%

Gene	Variant effect	aHUS					EVS					C3G			
		RVs	RVs cases	All cases	Frequency (%)	P^a	RVs	RVs cases	All cases	Frequency (%)	P^b	RVs	RVs cases	All cases	Frequency (%)
<i>CFH</i>	Protein-altering	193	411	3128	13.14	<0.0001 ^c	55	55	6503	0.08	<0.0001 ^c	22	23	443	5.19
<i>C3</i>	Protein-altering	64	177	2455	7.21	<0.0001 ^c	71	71	6503	1.09	<0.0001 ^c	24	26	379	6.86
<i>CD46</i>	Protein-altering	74	219	2942	7.44	<0.0001 ^c	95	95	6502	1.46	0.18	0	0	406	0.00
<i>CFI</i>	Protein-altering	68	98	2923	3.35	<0.0001 ^c	29	29	6494	0.45	0.48	2	2	408	0.49
<i>DGKE</i>	Protein-altering	19	21	703	2.99	<0.0001 ^c	24	24	6501	0.37	0.45	0	0	127	0.00
<i>THBD</i>	Protein-altering	6	7	1328	0.53	0.28	19	19	6457	0.29	0.0052	3	4	264	1.52
<i>CFB</i>	Protein-altering	15	22	2457	0.90	<0.0001 ^c	5	5	6503	0.08	<0.0001 ^c	6	5	379	1.32
<i>PLG</i>	Protein-altering	2	2	286	0.70	0.89	30	30	6503	0.46	0.16	2	2	104	1.92
<i>CFHR5</i>	Protein-altering	2	2	317	0.63	0.94	39	39	6502	0.60	0.64	1	1	104	0.96
<i>CFH</i>	Non-truncating ^e	129	318	3128	10.17	<0.0001 ^c	55	55	6503	0.85	<0.0001 ^c	17	19	443	4.29
<i>C3</i>	Non-truncating	59	171	2455	6.97	<0.0001 ^c	70	70	6503	1.08	<0.0001 ^c	20	21	379	5.54
<i>CFI</i>	Non-truncating	55	83	2923	2.84	<0.0001 ^c	26	26	6503	0.40	0.78	2	2	408	0.49
<i>CD46</i>	Non-truncating	43	99	2942	3.37	<0.0001 ^c	92	92	6502	1.41	0.03	0	0	406	0.00
<i>DGKE</i>	Non-truncating	7	9	703	1.28	<0.002 ^c	24	24	6501	0.37	0.49	0	0	127	0.00
<i>CFHR5</i>	Non-truncating	1	1	317	0.32	0.86	36	36	6503	0.55	0.58	1	1	104	0.96
<i>THBD</i>	Non-truncating	6	7	1328	0.53	0.19	17	17	6467	0.26	0.0026	3	4	264	1.52
<i>CFB</i>	Non-truncating	15	22	2457	0.90	<0.0001 ^c	4	4	6503	0.06	<0.0001 ^c	5	4	379	1.06
<i>PLG</i>	Non-truncating	2	2	286	0.70	0.8934	30	30	6503	0.46	0.1561	2	2	104	1.92
<i>CD46</i>	Truncating ^f	31	125	2942	4.25	<0.0001 ^c	3	3	6503	0.05	0.50	0	0	406	0.00
<i>CFH</i>	Truncating	64	98	3128	3.13	<0.0001 ^c	0	0	6503	0	<0.0001 ^c	5	5	443	1.13
<i>DGKE</i>	Truncating	12	18	703	2.56	<0.0001 ^c	0	0	6501	0	0.50	0	0	127	0.00
<i>CFI</i>	Truncating	13	16	2923	0.55	<0.0001 ^c	3	3	6422	0.05	0.50	0	0	408	0.00
<i>C3</i>	Truncating	5	6	2455	0.24	0.0024^c	1	1	6503	0.02	<0.0001 ^c	4	5	379	1.32
<i>THBD</i>	Truncating	0	0	1328	0.00	0.52	2	2	6373	0.03	0.77	0	0	264	0.00
<i>CFB</i>	Truncating	0	0	2457	0.00	0.54	1	1	6503	0.02	0.23	1	1	379	0.26
<i>PLG</i>	Truncating	0	0	286	0.00	0.84	0	0	6503	0	0.90	0	0	104	0.00
<i>CFHR5</i>	Truncating	1	1	317	0.32	0.46	3	3	6502	0.05	0.83	0	0	104	0.00

^a p value after the Chi-Square test (χ^2) with Yates' correction (p<0.05 after Bonferroni correction) for aHUS and EVS

^b p value after the Chi-Square test (χ^2) with Yates' correction (p<0.05 after Bonferroni correction) for C3G and EVS

^c Denotes aHUS or C3G dataset frequency significantly greater than ExAC frequency

^d Denotes EVS frequency significantly greater than aHUS or C3G dataset frequency

^e Defined as missense or in-frame variant

^f Defined as nonsense, frameshift, or splice acceptor/donor variant

Appendix IV. The total missense rare variant allele frequency for each protein domain, normalised by the proportion of domain residues (continues overleaf)

Protein	Domain	Residues (% of total in protein)	aHUS only			C3G only			RVs common to both aHUS and C3G				
			Missense RVs ^a	Sum AF ^b (%)	Sum AF/ residue proportion ^c	Missense RVs	Sum AF (%)	Sum AF/ residue proportion	Missense RVs	aHUS		C3G	
										Sum AF (%)	Sum AF/ residue proportion	Sum AF (%)	Sum AF/ residue proportion
C3	Signal peptide	22 (1.3)	2	0.08	6.1	0	-	-	0	-	-	-	-
C3	MG1	103 (6.2)	4	0.35	5.6	0	-	-	1	0.02	0.3	0.13	2.1
C3	MG2	103 (6.2)	10	1.34	21.7	1	0.13	2.1	0	-	-	-	-
C3	MG3	122 (7.3)	1	0.02	0.3	2	0.26	3.6	0	-	-	-	-
C3	MG4	98 (5.9)	0	-	-	1	0.13	2.2	0	-	-	-	-
C3	MG5	108 (6.5)	3	0.06	0.9	3	0.40	6.1	0	-	-	-	-
C3	MG6b	42 (2.5)	4	0.47	18.5	1	0.26	10.4	0	-	-	-	-
C3	LNK	65 (3.9)	0	-	-	1	0.26	6.8	2	0.22	5.7	0.40	10.1
C3	ANA	86 (5.2)	0	-	-	0	-	-	0	-	-	-	-
C3	aNT	17 (1)	0	-	-	0	-	-	0	-	-	-	-
C3	MG6a	59 (3.5)	1	0.06	1.7	2	0.26	7.4	0	-	-	-	-
C3	MG7	107 (6.4)	0	-	-	1	0.13	2.1	0	-	-	-	-
C3	CUBa	54 (3.2)	0	-	-	0	-	-	0	-	-	-	-
C3	TED	302 (18.2)	22	0.73	4.0	2	0.26	1.5	4	0.08	0.4	0.13	0.7
C3	CUBb	65 (3.9)	2	0.04	1.0	2	0.26	6.8	0	-	-	-	-
C3	MG8	139 (8.4)	7	0.14	1.7	2	0.26	3.2	3	0.06	0.7	0.13	1.6
C3	Anchor	22 (1.3)	1	0.02	1.5	0	-	-	0	-	-	-	-
C3	C345C	148 (8.9)	0	-	-	1	0.13	1.5	0	-	-	-	-
MCP	Signal peptide	34 (8.9)	3	0.10	1.1	0	-	-	0	-	-	-	-
MCP	SCR1	62 (16.2)	3	0.31	1.9	1	0.12	0.8	0	-	-	-	-
MCP	SCR2	63 (16.4)	10	0.31	1.9	0	-	-	0	-	-	-	-
MCP	SCR3	66 (17.2)	13	0.49	2.9	0	-	-	0	-	-	-	-
MCP	SCR4	60 (15.7)	11	0.37	2.4	0	-	-	0	-	-	-	-
MCP	Linker	59 (15.4)	4	0.09	0.6	0	-	-	0	-	-	-	-
MCP	TM	23 (6)	0	-	-	0	-	-	0	-	-	-	-
MCP	CT	16 (4.2)	1	0.07	1.6	0	-	-	0	-	-	-	-
FB	Signal peptide	25 (3.3)	0	-	-	0	-	-	0	-	-	-	-
FB	SCR1	66 (8.8)	0	-	-	0	-	-	0	-	-	-	-
FB	SCR2	60 (8)	1	0.02	0.3	1	0.13	1.7	0	-	-	-	-
FB	SCR3	58 (7.7)	0	-	-	0	-	-	0	-	-	-	-
FB	aL	49 (6.5)	0	-	-	0	-	-	5	0.10	1.5	0.26	4.0
FB	VWFA	200 (26.6)	13	0.43	1.6	2	0.26	1.0	0	-	-	-	-
FB	SP	295 (39.2)	4	0.08	0.2	3	0.53	1.3	0	-	-	-	-
FH	Signal peptide	18 (1.6)	2	0.03	2.0	1	0.11	7.1	0	-	-	-	-
FH	SCR1	60 (5.3)	8	0.29	5.5	1	0.11	2.1	1	0.02	0.3	0.11	2.1
FH	SCR2	57 (5)	1	0.02	0.3	4	0.57	11.3	0	-	-	-	-
FH	SCR3	60 (5.3)	8	0.21	3.9	2	0.23	4.3	1	0.02	0.3	0.11	2.1

Appendix IV. (continued) The total missense rare variant allele frequency for each protein domain, normalised by the proportion of domain residues (also continues overleaf)

Protein	Domain	Residues (% of total in protein)	aHUS only			C3G only			RVs common to both aHUS and C3G				
			Missense RVs	Sum AF ^b (%)	Sum AF/ residue proportion ^c	Missense RVs	Sum AF ^b (%)	Sum AF/ residue proportion ^c	aHUS		C3G		
									Missense RVs	Sum AF (%)	Sum AF/ residue proportion	Sum AF (%)	Sum AF/ residue proportion
FH	SCR4	53 (4.7)	6	0.11	2.4	0	-	-	0	-	-	-	-
FH	SCR5	54 (4.7)	1	0.02	0.3	0	-	-	0	-	-	-	-
FH	SCR6	61 (5.4)	4	0.06	1.2	0	-	-	0	-	-	-	-
FH	SCR7	54 (4.7)	3	0.14	3	0	-	-	1	0.02	0.3	0.23	4.8
FH	SCR8	58 (5.1)	1	0.02	0.3	0	-	-	0	-	-	-	-
FH	SCR9	56 (4.9)	6	0.27	5.5	0	-	-	0	-	-	-	-
FH	SCR10	55 (4.8)	5	0.1	2	0	-	-	1	0.06	1.3	0.23	4.7
FH	SCR11	55 (4.8)	6	0.14	3	0	-	-	0	-	-	-	-
FH	SCR12	54 (4.7)	1	0.02	0.3	0	-	-	0	-	-	-	-
FH	SCR13	51 (4.5)	1	0.02	0.4	0	-	-	0	-	-	-	-
FH	SCR14	53 (4.7)	8	0.19	4.1	0	-	-	1	0.02	0.3	0.11	2.4
FH	SCR15	57 (5)	7	0.13	2.6	1	0.11	2.3	1	0.02	0.3	0.11	2.3
FH	SCR16	54 (4.7)	3	0.10	2.0	0	-	-	0	-	-	-	-
FH	SCR17	55 (4.8)	3	0.06	1.3	0	-	-	0	-	-	-	-
FH	SCR18	55 (4.8)	6	0.30	6.3	2	0.33	4.6	0	-	-	-	-
FH	SCR19	55 (4.8)	8	0.21	4.3	0	-	-	0	-	-	-	-
FH	SCR20	62 (5.5)	31	2.03	37.2	0	-	-	2	0.38	7.1	0.23	4.2
FHR5	Signal peptide	19 (3.3)	0	-	-	0	-	-	0	-	-	-	-
FHR5	SCR1	61 (10.7)	0	-	-	0	-	-	0	-	-	-	-
FHR5	SCR2	58 (10.2)	0	-	-	0	-	-	0	-	-	-	-
FHR5	SCR3	59 (10.4)	0	-	-	0	-	-	0	-	-	-	-
FHR5	SCR4	59 (10.4)	0	-	-	0	-	-	0	-	-	-	-
FHR5	SCR5	58 (10.2)	0	-	-	0	-	-	0	-	-	-	-
FHR5	SCR6	55 (9.7)	0	-	-	0	-	-	0	-	-	-	-
FHR5	SCR7	58 (10.2)	1	0.16	1.6	1	0.48	4.7	0	-	-	-	-
FHR5	SCR8	59 (10.4)	1	0.16	1.5	0	-	-	0	-	-	-	-
FHR5	SCR9	63 (11.1)	0	-	-	0	-	-	0	-	-	-	-
FI	Signal peptide	18 (3.1)	0	-	-	0	-	-	0	-	-	-	-
FI	Linker 1	23 (4)	1	0.02	0.4	0	-	-	0	-	-	-	-
FI	FIMAC	66 (11.4)	9	0.24	2.1	1	0.12	1.1	0	-	-	-	-
FI	SRCR	103 (17.8)	13	0.68	3.8	0	-	-	0	-	-	-	-
FI	LDLR 1	44 (7.6)	6	0.12	1.6	0	-	-	0	-	-	-	-
FI	LDLR 2	36 (6.2)	2	0.07	1.1	0	-	-	0	-	-	-	-
FI	Linker 2	46 (7.9)	4	0.17	2.2	0	-	-	1	0.02	0.2	0.12	1.7
FI	SP	244 (43)	28	1.09	2.5	0	-	-	1	0.05	0.1	0.12	0.3
DGKE	Signal peptide	34 (6)	1	0.07	1.2	0	-	-	0	-	-	-	-
DGKE	DAGKa	155 (27.3)	0	-	-	1	-	-	0	-	-	-	-

Appendix IV. (continued) The total missense rare variant allele frequency for each protein domain, normalised by the proportion of domain residues

Protein	Domain	Residues (% of total in protein)	aHUS only			C3G only			RVs common to both aHUS and C3G				
			Missense RVs	Sum AF ^b (%)	Sum AF/ residue proportion ^c	Missense RVs	Sum AF ^b (%)	Sum AF/ residue proportion ^c	aHUS		C3G		
									Missense RVs	Sum AF (%)	Sum AF/ residue proportion	Sum AF (%)	Sum AF/ residue proportion
DGKE	DAGKc	141 (24.9)	3	0.28	1.1	0	-	-	0	-	-	-	-
DGKE	C1	51 (9)	2	0.14	1.6	1	0.39	4.4	0	-	-	-	-
DGKE	C2	53 (9.3)	0	-	-	0	-	-	0	-	-	-	-
DGKE	Linker	43 (7.6)	1	0.07	0.9	0	-	-	0	-	-	-	-
PLG	Signal peptide	19 (2.3)	0	-	-	0	-	-	0	-	-	-	-
PLG	PAN	79 (9.8)	0	-	-	0	-	-	0	-	-	-	-
PLG	Kringle 1	79 (9.8)	1	0.18	1.8	2	0.96	9.9	0	-	-	-	-
PLG	Kringle 2	79 (9.8)	1	0.18	1.8	0	-	-	0	-	-	-	-
PLG	Kringle 3	78 (9.6)	0	-	-	0	-	-	0	-	-	-	-
PLG	Kringle 4	78 (9.6)	0	-	-	0	-	-	0	-	-	-	-
PLG	Kringle 5	80 (9.9)	1	0.18	1.8	0	-	-	0	-	-	-	-
PLG	Peptidase S1	228 (28.1)	0	-	-	0	-	-	0	-	-	-	-
PLG	Linker	25 (3.1)	1	0.18	5.7	0	-	-	0	-	-	-	-
THBD	C-type lectin	139 (24.2)	7	0.57	2.3	0	-	-	0	-	-	-	-
THBD	EGF-like 6	41 (7.1)	1	0.04	0.5	0	-	-	2	0.34	4.8	0.57	8.0
THBD	EGF-like 1	41 (7.1)	0	-	-	0	-	-	0	-	-	-	-
THBD	EGF-like 3	39 (6.8)	0	-	-	0	-	-	0	-	-	-	-
THBD	EGF-like 2	41 (7.1)	0	-	-	1	0.19	2.7	0	-	-	-	-
THBD	EGF-like 5	37 (6.4)	0	-	-	0	-	-	0	-	-	-	-
THBD	EGF-like 4	41 (7.1)	0	-	-	0	-	-	0	-	-	-	-
THBD	Integrin binding	33 (5.7)	0	-	-	1	0.38	6.6	1	0.15	2.6	0.19	3.3

^a Number of missense RVs classified as ‘pathogenic’, likely pathogenic’, or ‘uncertain significance’

^b The sum of missense RV allele frequencies for each protein domain

^c The sum of missense RV allele frequencies divided by the proportion of protein residues attributed to the protein domain